Cognition 120 (2011) 268–291

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Investigating joint attention mechanisms through spoken human–robot interaction

Maria Staudte *, Matthew W. Crocker

Department of Computational Linguistics Campus, Saarland University, 66123 Saarbruecken, Germany

ARTICLE INFO

Article history:
Received 18 March 2010
Revised 27 April 2011
Accepted 11 May 2011
Available online 12 June 2011

Keywords:
Utterance comprehension
Referential gaze
Joint attention
Human–robot interaction
Referential intention
Gaze
Situated language processing
Reference resolution

ABSTRACT

Referential gaze during situated language production and comprehension is tightly coupled with the unfolding speech stream (Griffin, 2001; Meyer, Sleiderink, & Levelt, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In a shared environment, utterance comprehension may further be facilitated when the listener can exploit the speaker's focus of (visual) attention to anticipate, ground, and disambiguate spoken references. To investigate the dynamics of such gaze-following and its influence on utterance comprehension in a controlled manner, we use a human–robot interaction setting. Specifically, we hypothesize that referential gaze is interpreted as a cue to the speaker's referential intentions which facilitates or disrupts reference resolution. Moreover, the use of a dynamic and yet extremely controlled gaze cue enables us to shed light on the simultaneous and incremental integration of the unfolding speech and gaze movement.

We report evidence from two eye-tracking experiments in which participants saw videos of a robot looking at and describing objects in a scene. The results reveal a quantified benefit-disruption spectrum of gaze on utterance comprehension and, further, show that gaze is used, even during the initial movement phase, to restrict the spatial domain of potential referents. These findings more broadly suggest that people treat artificial agents similar to human agents and, thus, validate such a setting for further explorations of joint attention mechanisms.

1. Introduction

According to an old and widespread proverb, the eyes are windows to the soul. The validity of this statement has, at least to some extent, been supported by a large body of previous psychological and psycholinguistic research. Baron-Cohen, Baldwin, and Crowson (1997) state, for instance, that *"mental states can (...) be read from direction of gaze. These include desire, refer, and goal* (Baron-Cohen, Campbell, Karmiloff-Smith, Grant, & Walker, 1995). *That is, our natural reading of gaze directed at a specific object is in terms of a person's volitional states. This should come as no surprise, since we tend to look at what we want, and to what we are referring, and at what we are about to act upon."* (p. 312).

Undoubtedly, the primary function of directing gaze is related to the act of *seeing*. To fixate something or someone lets us inspect that object or person in greater detail. Additionally, gaze in communication reflects numerous different processes and responds to many cues. It conveys, for instance, information about emotions, goals or desires (Argyle & Dean, 1965; Baron-Cohen, Wheelwright, & Jolliffe, 1997; Dovidio & Ellyson, 1982). In addition to these meta-linguistic functions, gaze can also reflect information that is directly linked to the content of a spoken utterance. A deictic expression accompanied by a glance towards a certain object may be a valid and comprehensible reference for a listener in face-to-face communication (Clark & Krych, 2004). Thus, a listener seems to be able to link

* Corresponding author. Tel.: +49 (0)681 302 6552.
E-mail address: masta@coli.uni-saarland.de (M. Staudte).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.05.005

the spoken reference to the object which is in focus of the speaker's visual attention.

Previously, gaze has indeed been widely studied as an indicator for overt visual attention during language processing and it was shown that where we look is closely related to what we say and understand. Studies have revealed, for instance, that speakers look at entities roughly 800–1000 ms before mentioning them (Griffin, 2001; Meyer et al., 1998), while listeners inspect objects as soon as 200–400 ms after the onset of the corresponding referential noun (Allopenna, Magnuson, & Tanenhaus, 1998; Tanenhaus et al., 1995). This shows that eye gaze during situated language production and comprehension is tightly coupled with the unfolding speech stream. In face-to-face communication, the speaker's gaze to mentioned objects in a shared environment also provides the listener with a visual cue as to the speaker's focus of (visual) attention (Flom, Lee, & Muir, 2007). Following this cue (among other cues) in order to attend to the same things as the interlocutor has been dubbed *joint attention* (e.g., Emery, 2000). By revealing a speaker's focus of visual attention, such gaze cues potentially offer the listener valuable information to ground and sometimes disambiguate referring expressions, to hypothesize about the speaker's communicative intentions and goals and, thus, to facilitate comprehension (e.g., Clark & Krych, 2004; Hanna & Brennan, 2007).

The goal of this paper is to systematically investigate how speaker gaze influences real-time listener comprehension. Specifically, we examine how listeners exploit gaze cues during an unfolding sentence and whether this facilitates situated comprehension. The following three hypotheses address the main questions we aim to answer:

1. We hypothesize, firstly, that referential speaker gaze is in fact beneficial for utterance comprehension and that this benefit can be quantified (Experiment 1).
2. Secondly, we hypothesize that such an effect of referential speaker gaze is due to the intentional status of gaze (see Becchio, Bertone, & Castiello, 2008) rather than a purely reflexive shift of visual attention (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999, Experiment 2).
3. We thirdly hypothesize that (even reflexive) gaze-following may be split up into the first phase, in which listeners potentially follow already the gaze movement (which is comparable to the human head movement often accompanying gaze), and the second (fixation) phase, which has traditionally been studied by means of static gaze cues (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999, Experiment 1).

Examining these issues in human–human interaction is problematic, however. Not only is speaker gaze likely to be inconsistent and "noisy", but it is also difficult to plausibly elicit incongruent speaker gaze behavior which is necessary to quantify any benefits of natural gaze. The solution we adopt is to use an artificial robot agent as the speaker. This provides greater uniformity and control of the gaze–speech synchronization while the fallibility of robots makes non-natural gaze behavior more plausible. The broad control over all gaze parameters such as direction, duration, or onset can further be exploited to examine in more detail how closely listeners attend to speaker gaze and whether it is followed as early as gaze movement onset.

It is of course an open question in itself whether joint-attention-like behavior is unique to human–human interaction – possibly hinging on common biological and cognitive mechanisms – or whether such gaze cues play a similar role in human–robot interaction. Previous research has shown, however, that people readily attribute intentional states and personality traits to non-humans as well, like animals or artificial agents such as robots (see e.g. Nass & Moon, 2000, Kiesler, Powers, Fussell, & Torrey, 2008, for overviews). We therefore also examine the hypothesis that joint attention and its effects on communication are an important component of human–robot interaction as well. That is, when a robot produces human-like aligned gaze and speech behavior such that its gaze indicates which object is in visual focus, we predict that listeners will jointly attend to that object. Such a finding would therefore strengthen the use of human–robot interaction scenarios to conduct controlled studies of mechanisms and phenomena that typically occur when people establish joint attention with other people, thus, extending the theoretical value of these studies to human–human interaction.

Supporting evidence for these hypotheses is provided by two eye-tracking experiments exploiting different tasks in a human–robot interaction (HRI) scenario. Additionally, these results shed light on the dynamics of gaze and speech integration and its effects on utterance interpretation.

1.1. The coupling of gaze and language

Since language is often vague and ambiguous, additional non-verbal cues supporting and augmenting the conveyed message or the retrieval of information are potentially useful in face-to-face communication. While cues like pointing generally complement spoken language and are often useful for grounding and disambiguating an utterance in the scene (Bangerter, 2004), gaze appears to have a special status among such non-verbal cues: Gaze is permanently available since people constantly use and move their eyes even when their gaze is not related to language production or comprehension and even when they do not intend to signal anything at all. Further, gaze is highly diverse in its expressiveness conveying various mental states (as above see, e.g., Adams & Kleck, 2003; Baron-Cohen, Wheelwright, et al., 1997; Dovidio & Ellyson, 1982).

Certainly, the desire or need to 'see' is a major initiator for gaze movement and frequently elicited by cues in the visual scene itself (Henderson, 2003), sometimes including also a speaker's gaze. Furthermore, a close coupling of language and (referential) gaze has been established in a number of studies, showing that where people look is often driven by what they hear or say (Allopenna et al., 1998; Altmann & Kamide, 1999; Altmann & Kamide, 2004; Griffin & Bock, 2000; Knoeferle, Crocker, Pickering, & Scheepers, 2005; Meyer et al., 1998; Tanenhaus et al., 1995). It is

possibly because of this systematic and automatic coupling between speech and gaze that listeners are able to interpret speakers' eye-movements on-line as visual references.

Whether, and how, the close alignment of visuo-linguistic processes helps listeners to comprehend utterance content, is subject to ongoing research (Crocker, Knoeferle, & Mayberry, 2010). Previous studies suggest that people do indeed monitor and use each other's gaze and speech in face-to-face communication to rapidly ground and resolve spoken utterances with respect to a common environment (Clark & Krych, 2004; Moore & Dunham, 1995; Tomasello & Carpenter, 2007). That is, where a speaker looks may constrain the domain of interpretation for the listener (Hanna & Brennan, 2007) and where a listener looks may tell the speaker that she has misunderstood such that the speaker can decide to repeat or to further specify a referring expression (Clark & Krych, 2004). However, the precise *temporal* and *causal* relationship between gaze cues and concurrent language processing has, to our knowledge, not been fully explored.

1.2. Joint and shared attention

Listeners may use speaker gaze as a timely cue to utterance content, possibly *because* of the tight coupling of gaze and speech mentioned above. In order to understand why and how people use each other's gaze as referential cues, the notion of visual attention is essential. It helps to establish and understand the connection between eye gaze and its referents in the external world. Allocation of visual attention allows more detailed inspection of one aspect in the environment (*selectivity*) while limiting processing of other (visual) information (*capacity limitation*, Bundesen, 1990; Desimone & Duncan, 1995). That is, an entity that is being *looked* at is typically in the focus of visual attention, allowing investigation of the entity's visual features in greater detail. Consequently, following the interlocutor's gaze typically reveals information about what she is or has been visually attending to and may result in two people jointly attending to the entity in question (Moore & Dunham, 1995).

Following Emery (2000), we consider *joint attention* to occur when an individual follows another individual's gaze, for instance, to mutually attend to an entity while possibly inferring her referential intentions. The term 'referential intention' is based on the definition of intention put forward in Tomasello, Carpenter, Call, Behne, and Moll (2005): *"So the organism has the goal "that X be the case" and the intention "to do A" in pursuit of that goal."* (p. 677), and refers specifically to the speaker's intention to communicate a message about certain entities in the environment which is done in form of a spoken utterance and possibly any accompanying non-verbal cues. Thus, being able to jointly attend to an object and inferring a referential intention presupposes that the gazer actually has attentional states such that the gaze-follower has reason to consider the looked-at entity as relevant for communication purposes. Further, the term *shared attention* is used to refer to a phenomenon which presupposes a higher level of interactivity and relates to what Tomasello et al. (2005) calls *shared intentionality*: *"Collaborative interactions in which participants have a shared goal (...) and coordinated action roles for pursuing that shared goal."* (p. 677). That is, shared attention implies that one person *intentionally* directs another person's gaze to an object in order to coordinate action and perception for achieving a mutual goal or just to share the experience (see also Emery, 2000).

Notably, what we call shared attention has previously also been named joint attention (Kaplan & Hafner, 2006; Tomasello & Carpenter, 2007). Similarly, it has been described as a state that requires that the *"goal of each agent is to attend to the same aspect of the environment"* (Kaplan & Hafner, 2006, p. 144) and that "both agents are aware of this coordination of 'perspectives' towards the world" (Kaplan & Hafner, 2006, p. 145). However, we adopt a more fine-grained categorization and distinguish joint and shared attention in order to account for the limited interactivity in our experimental design where shared attention is essentially infeasible.

1.3. Gaze in human–computer interaction

Using an artificial agent to investigate joint attention mechanisms presupposes that people consider such an agent similar enough to themselves in order to assign attentional and intentional states to it during interaction. To confirm this – beyond simply showing that people readily assign human traits to artificial agents in general (Nass & Moon, 2000) – we address the question of whether people try at all to align their attention with the agent in a natural, joint attention-like way, and whether this affects their utterance processing.

Despite the generally growing interest in human–computer/human–robot interaction (HCI/HRI) to incorporate natural gaze mechanisms, the fine-grained effects of closely aligned referential gaze and speech described above have not been systematically investigated. Rather, previous work on gaze in HCI/HRI has concentrated largely on the general appearance of the agent and what competencies and characteristics people intuitively ascribe to agents featuring certain gaze behaviors. Kanda, Ishiguro, and Ishida (2001), for instance, equipped their robot with very basic gaze movements and observed that people generally found the interaction more enjoyable than when the robot showed no gaze movements. Thus, robot gaze can, on one hand, improve agreeableness of HRI. On the other hand, robot gaze can be dysfunctional and disturb smooth interaction: Despite increasing general agreeableness, the robot's crude gaze movements resulted in a lower performance judgement revealed by a post-experiment questionnaire. Similarly, Sidner, Lee, Kidd, Lesh, and Rich (2005) found that participants judged the robot they had to interact with to be less 'reliable' when it showed gaze (or head) movement.

Cassell, Torres, and Prevost (1999) took a different approach to implementing more natural gaze behavior, on the one hand, and to explore the utility of mutual gaze in general on the other hand. According to the psychology literature, mutual gaze (i.e., looking at each other) is a signal that is used to coordinate turn-taking in a conversation (Duncan, 1972; Kendon, 1967). Cassell and her colleagues hypothesized that gaze also correlates with information

structure of the discourse, that is, theme (what is known, links the utterance to previous discourse) and rheme (new information) of an utterance. An initial study confirmed the correlation of speaker gaze towards and away from the listener with both turn-taking and information structure (Cassell, Torres, et al., 1999). The authors thus developed and implemented a heuristic for gaze production as a realization of turn-taking cues and equipped an embodied conversational agent with such a behavior (Cassell, Bickmore, et al., 1999; Cassell & Thórisson, 1999). Post-experiment questionnaires from a user study involving such an agent revealed that users felt the agent to be more helpful, lifelike, and smooth when it showed this nonverbal conversational behavior. More recently, Mutlu and colleagues implemented the initial probabilistic algorithm suggested by Cassell, Torres, et al. (1999) drawing on both turn-taking and information structure effects. This implementation was used and evaluated on a storytelling humanoid robot (Mutlu, Hodgins, & Forlizzi, 2006). Results showed that participants who were looked at by the robot more often performed better on the recall task, suggesting that people attend closer to the robot when being looked at more frequently (at the appropriate occasions during discourse).

The mentioned studies in HCI suggest that gaze in one way or the other affects the *impression* a person or agent makes. Psycholinguistic evidence reported above shows that gaze can additionally provide concrete information that helps to quickly link the accompanying utterance to the world and guide attention accordingly. There has been limited research in HCI, however, that explores the use of gaze as a visual modality which augments speech and elicits joint attention with an artificial agent and which may be used to ground and disambiguate spoken references. The work conducted by Breazeal and colleagues with the robot *Leonardo* (Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005) has made relevant contributions to this field of research. It was shown, for instance, that implicit robot behavior, like gaze shifts, head nods, and other gestures, was used by human interlocutors to solve a collaboration task faster. In one such study, participants were asked to interact with *Leonardo* and make it switch on buttons that were located in front of it. Results revealed that the robot's implicit, non-verbal information helped people to detect errors in the robot's performance and, consequently, to repair them. Not surprisingly, task completion time was considerably shorter in the implicit condition than when no such non-verbal information was available. In particular, such facilitation effects were observed when misunderstandings and other errors occurred during the conversation which needed to be repaired. Breazeal et al. (2005) concluded that Leonardo's gaze constituted a "window to its visual awareness" and that people perceive the robot's gaze as signaling to "share attention" (Breazeal et al., 2005, p. 714)

The findings on the role of gaze in HCI/HRI, as reported in this section, provide promising support for the hypothesis that people do seek to establish joint attention with an artificial agent using gaze cues. However, these results are largely based on subjective measures taken off-line and effects on participants' performance may simply be due to agent gaze behavior engaging participants at a very general level. In the following section, we draw on the findings from the various research areas reported above to motivate our own investigation of the role of referential speaker gaze in situated human–robot interaction. Specifically, we set out to examine the on-line influence of (robot) speaker gaze, *simulating the gaze and speech alignment of a human speaker*, on the listener's visual attention and utterance comprehension.

1.4. Does coupling of robot gaze and language elicit joint attention?

The psycholinguistic findings mentioned earlier can be summarized in the following scenario: Two interlocutors (A and B) are talking about objects in their shared view. According to the production literature, A might describe two objects to B in the following way. While saying "The cylinder is taller than the pyramid that is pink.", A looks at the cylinder approximately 800–1000 ms before uttering "cylinder" and then looking at the pyramid approximately 800–1000 ms before the critical reference to the pyramid (Griffin & Bock, 2000). The comprehension literature using the Visual World Paradigm further established that as listener B is processing A's utterance, she typically looks at the cylinder and pyramid around 200–300 ms after A started saying "cylinder" and "pyramid", respectively (Cooper, 1974; Tanenhaus et al., 1995). If additionally A and B can see each other, joint visual attention may be established throughout the interaction even earlier compared to when only speech is available. Listener B can follow A's gaze towards the pyramid right away and anticipate A's mentioning of the mug (Hanna & Brennan, 2007). The time span between A's and B's gaze towards a mentioned object is shortened dramatically and B may rapidly ground A's reference to the pyramid in their common view.

To investigate whether listeners similarly attend to and use speech-mediated *robot* gaze, we created a setup where listeners (cf. B above) viewed a robot speaker (see Fig. 1, representing A from above) that described objects in a scene while looking at those objects. Participants were eye-tracked while observing these videos. They were additionally asked to quickly determine the 'correctness' of the given statement with respect to the scene by pressing a button (Experiment 1), or to correct false statements orally (Experiment 2). Thus, we consider three dependent measures: Listeners' eye-movements in the scene as an on-line measure during comprehension, as well as response times and correction statements as off-line measures to indirectly asses comprehension time and interpretation, respectively.

Although it might be argued that such a video-based presentation mode does not allow true interaction, it has been shown that a video-based scenario without true interaction yields similar results to a live-scenario and can be considered to similarly provide valuable insights into participants' perception and opinions (Woods, Walters, Koay, & Dautenhahn, 2006). Further, the subjective perception of remote versus co-located virtual agents and robots has been studied. On the one hand, results from a user study conducted by Kiesler et al. (2008) suggest that people have different expectations and impressions of an

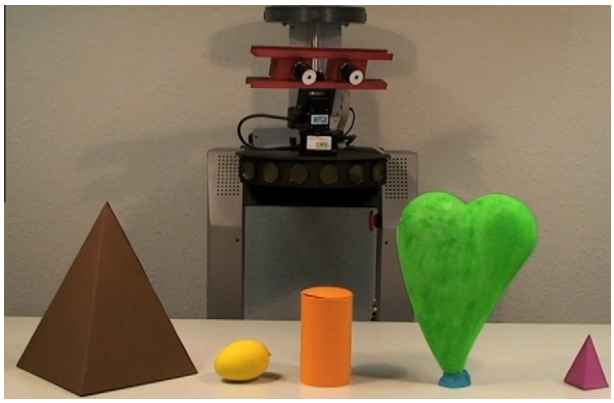


Fig. 1. Robot speaker. Its head and gaze direction is realized by the stereo-camera mounted on a pan-tilt-unit. A German sample utterance *"Der Zylinder ist grösser als die Pyramide, die pink ist."* (English: *"The cylinder is taller than the pyramid that is pink."*) would be accompanied by gaze, e.g., towards the cylinder and then towards the small, pink pyramid.

agent versus a robot: Mistakes and errors are potentially more acceptable and less irritating when communicating with a robot and, simultaneously, the robot was perceived as more life-like, having more positive personality traits and being liked better. On the other hand, results from the same study showed that the presentation mode of the robot or agent (co-present versus remote, i.e., recorded and projected onto a screen) did not greatly influence participants' impressions.

We consider these findings to support our decision to employ a robot for studying referential gaze behavior and to indicate that remote, video-based presentation should not substantially affect perception thereof. At the same time, we are aware that video-based presentation limits interactivity and therefore we carefully employ the concept of joint attention – in contrast to shared attention – in this context. It is arguably the case that a speaker has a referential intention and acts upon it when being video-taped. When another individual views this behavior (even if at a later stage on video), we expect this individual to draw on natural comprehension mechanisms. Thus, a video-based experimental design may nevertheless allow for the investigation of some phenomena that are typically involved during joint and shared attention as well.

Our video stimuli were manipulated with respect to congruency of both the linguistic and visual (gaze) referential cues, firstly, to examine whether people infer referential intentions from human-like aligned gaze (as in Fig. 1) and whether that facilitates utterance comprehension. Secondly, we sought to investigate how people deal with gaze cues that are incongruent or misaligned with the spoken utterance but which are not necessary to understand the utterance (Fig. 2 shows incongruent robot gaze to the brown, tall pyramid during the same utterance as in Fig. 1). Such situations occur, for instance, when misunderstandings lead to the use of inappropriate objects names, or errors in spatial memory elicit a glance towards another object – both in human–human or human–computer interaction. In the latter, incongruent multi-modal references (i.e., conflicting linguistic and visual cues) could easily be caused by an agent's "misprogrammed" gaze movements or errors in its object recognition. Furthermore, insights on how inappropriate co-occurrences of gaze and speech cues are resolved offer the potential to illuminate the nature of gaze influence as well as the integration process of information that is provided through different modalities such as language and vision.

Based on the findings about the use of speaker gaze reported from HHI and HRI investigations, we identify two possible stages of response mechanisms in our scenario which determine in what way speaker gaze influences situated utterance comprehension more generally.

1. The Visual Account: Listeners may follow robot gaze, possibly reflexively as observed in response to stylized gaze cues (and other symbolic cues such as arrows) in previous studies (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999). These so-called *exogenous* cueing effects found in those studies are typically short-lived. Yet, it is an interesting question to what extent the visual information obtained after such an attention shift may also affect further (visual and linguistic) processing. Indeed, recent findings involving exogenous visual cues have suggested that trial-initial focus of attention influences scene apprehension and the structure of produced scene descriptions (e.g., Gleitman, January, Nappa, & Trueswell, 2007; Tomlin, 1997).
2. The Intentional Account: If listeners treat the robot's camera movement as a type of eye gaze – that is, they accept it as a way of *seeing*, similar to human eyes and with similar functions – we predict that people will use robot gaze as an attentional cue. While gaze-following may still be reflexive, using gaze *additionally* as a cue that reflects the speaker's visual attention, and is associated with a referential intention, would result in joint attention (as indicated, for instance, by Becchio et al., 2008). We envisage this phenomenon as a "secondary" association between speaker gaze and her communicative goals, possibly *after* or *beyond* any initial, reflexive attention shifts. That is, human listeners may jointly attend to what the (robot) speaker attends to, thereby anticipating and grounding the next referent. Utterance comprehension may thus be affected

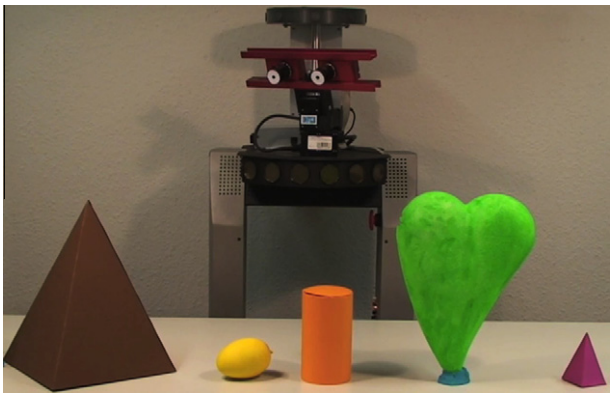


Fig. 2. In contrast to Fig. 1, the utterance *"The cylinder is taller than the pyramid that is pink."* is now accompanied by robot gaze to the cylinder and the tall, brown pyramid, resulting in an **incongruent** multi-modal reference.

on-line by speaker gaze in terms of *which* object is considered to be the referent and/or *how quickly* the reference is resolved.

The Intentional Account has been supported more recently by a number of studies (Bayliss, Paul, Cannon, & Tipper, 2006; Becchio et al., 2008; Castiello, 2003; Meltzoff, Brooks, Shon, & Rao, 2010; Vecera & Rizzo, 2006). Providing further supporting evidence in a human–robot interaction scenario would simultaneously strengthen the validity of our experimental design and the generalizability to human–human interaction. While there are several previous studies generally supporting the view that people indeed anthropomorphize and ascribe complex mental states to a robot (such as team spirit or the level of dominance or expertise, Nass & Moon, 2000, but also intentionality, Meltzoff et al., 2010), our experiments are of a different nature. On the one hand, our measures are on-line, revealing the temporal dynamics in greater detail. On the other hand, the manipulations in our studies consist in a simple visual cue, only roughly approximating human gaze behavior. Further, we would like to point out that while our gaze manipulation may be a simple cue in the context of HRI, it is still a dynamic cue that interacts with scene apprehension and simultaneous language processing. Thus, in the context of psycholinguistic studies on the interplay of visual and linguistic processing, such a complex and dynamic setting provides a novel way to explore the real-time integration of this multi-modal information. Our findings therefore address not only the hypothesis that people try to establish joint attention with a robot, but further reveal details of the dynamic integration of gaze and speech more generally.

Specifically, Experiment 1 investigated the first hypothesis, that speaker gaze indeed *facilitates* comprehension of the utterance while also showing that people follow the speaking robot's gaze, at least reflexively (Visual Account). Moreover, this experiment sheds light on the incremental use of gaze (movement and fixation) during an unfolding utterance (third hypothesis). Experiment 2 further served to replicate gaze and speech-following behavior under a correction task while investigating whether speaker gaze further influenced what people considered as the *intended* referent (second hypothesis). Results indeed support the Intentional Account, suggesting that listeners follow speaker gaze (possibly reflexively) and, more importantly, that they *use* the information about what is in visual focus to predict a referent.

2. Experiment 1

Experiment 1 examined whether referential speaker gaze is followed and used for reference resolution even though utterances could be validated without paying any attention to speaker gaze. Specifically, we investigated and quantified the actual benefit of speaker gaze, as assessed by comparing response times for sentence validity judgments when utterances were accompanied by referential gaze and when they were accompanied by neutral gaze. Observing a facilitatory effect of congruent referential speaker gaze would further indicate that listeners indeed establish a link between looked-at and mentioned objects.

When considering referential gaze and its utility for reference resolution, the question naturally arises whether referential gaze could also disrupt reference resolution if, for instance, it identified an entity other than the one referenced in the utterance. If people assigned attentional states to the robot such that they assumed that an object looked-at by the robot was likely the one it intended to mention, then incongruent referential gaze would be a misleading cue that would disrupt utterance comprehension.

Thus, to further investigate the role of speaker gaze, we manipulated the *Gaze Congruency* of speaker gaze as a potential cue for intended meaning as well as the *Validity* of the statements. Statements were either true or false, that is, the stated relationship between objects held or not, and the visual reference (established by robot gaze) was either congruent, incongruent or neutral with respect to the linguistic reference. Gaze was considered to be congruent (and helpful) when it was directed towards the same object that was going to be mentioned shortly afterwards (reference match, see also Fig. 1) while it was considered as incongruent when gaze was directed to an object different from the mentioned referent (mismatch, Fig. 2). In a

third Gaze Congruency level, robot gaze was neutral. The robot briefly looked down at the scene and back towards the camera – or a potential listener – before beginning to utter a scene description. The neutral gaze behavior provided a baseline condition in which listeners' visual attention was purely a response to the produced robot utterance and comprehension was uninformed by any joint attention mechanisms.

Finally, the temporal ambiguity in the utterance concerning the target reference, as given in Fig. 1, enabled us to look more closely at the real-time use of gaze cues by the listener, and whether the gaze movement could restrict the spatial domain of potential referents *prior* to the actual fixation.

The scene provided one referent for the cylinder (the "anchor") and two potential referents for the "target" noun (e.g., two pyramids of different sizes and colors), one of which the robot mentioned explicitly. One pyramid matched the description of the scene (*was shorter* than the cylinder) while the other did not (it was actually *taller* than the cylinder). Thus, which pyramid was finally mentioned depended on the sentence final color adjective and determined whether the statement was valid or not. The manipulation of both factors, Sentence Validity (true, false) and Gaze Congruency (congruent, neutral, incongruent), resulted in six conditions per item. Fig. 3 provides a sample scene as well as the set of all conditions that the corresponding example sentence appeared in.

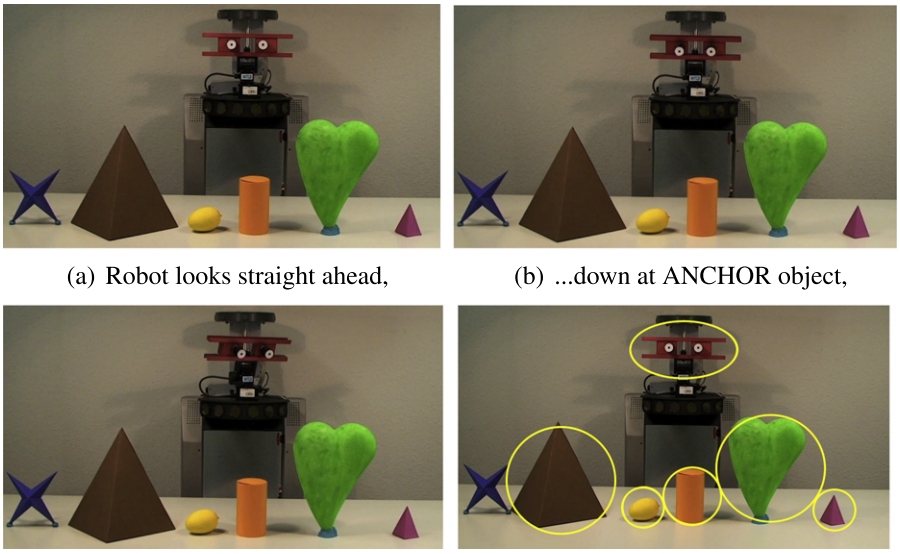
2.1. Method

2.1.1. Participants

Forty-eight native speakers of German, mainly students enrolled at Saarland University, took part in this study. All participants reported normal or corrected-to-normal vision. Most of them had no experience with robots.

2.1.2. Materials

A set of 24 items was used. Each item consisted of three different videos and two different sentences. Additionally we counterbalanced each item by reversing the comparative in the sentence predicate, for instance, from "taller" to "shorter", such that the target became the competitor and vice versa. We obtained a total of twelve videos per item while ensuring that target size, location and color were balanced. All versions showed the same scene and only differed with respect to where the robot looked and whether it verbally referred to the correct (target) object. Twelve different object shapes appeared twice each as target–competitor pairs to produce 24 items. For each shape, we created three different sizes (small, medium, large) and used small-large pairs as target–competitor pairs and the



(a) Robot looks straight ahead, (b) ...down at ANCHOR object, (c) ...at TARGET object, (d) ...and back up. (with marked regions of interest)

Robot Gaze	**True** Sentence: *"The cylinder is taller than the pyramid that is pink."*	**False** Sentence: *"The cylinder is taller than the pyramid that is brown."*
Congruent	looks to small pink pyramid	looks to big brown pyramid
Incongruent	looks to big brown pyramid	looks to small pink pyramid
Neutral	–	–

Fig. 3. Pictures illustrate a sample scene and robot gaze movements from Experiment 1. The manipulation of sentence validity and robot gaze results in the six conditions described below the pictures.

medium sized shape as anchor for another target–competitor pair. Moreover, each scene contained three additional distractors, one large and one small one, positioned to either side of the anchor. They served as potential competitors for partial utterances up to the comparative (e.g., "The pyramid is taller than"). The third distractor was typically small and positioned to the far left or far right of the scene.

Prior to the experiment, target–competitor pairs were pre-tested in order to make sure that their size and color differences were easily recognizable. We used a questionnaire that showed photographs of the original scenes excluding the robot. Twenty participants judged whether a given item sentence accurately described what was visible in the scene. For each scene, three sentences were given and only one of those contained a comparison between item objects (anchor and target/competitor). Overall, 50 % of the sentences were true and 50 % were false in order to avoid an acquiescence bias. A 7-level Likert scale from 1 (incorrect) to 7 (correct) allowed for a graded judgement of the sentences' validity. The results exhibit a mean deviation of 0.26 points from the optimal answer (1 and 7) which clearly showed that the comparisons between the distinct objects and their sizes were clear and easily assessable.

We created 1920×1080 resolution video-clips showing the PeopleBot[1] robot onto which a pan-tilt unit was mounted, carrying the stereo camera. Note, that head orientation and eye-gaze are identical for this robot. Further, the robot was positioned behind a table with a set of colored objects in front of it. After a two-second preview time, the video-clips each showed a sequence of camera-movements consecutively towards the anchor and the target/competitor objects. Simultaneously, a synthesized sentence such as given in Fig. 1 was played back. Sentences were in German and synthesized using the Mary TTS system (Schroeder & Trouvain, 2003). We overlaid the videos of the moving robot with the spoken stimulus sentences such that a robot fixation towards an object occurred one second prior to the onset of the referring noun, consistent with corresponding findings on alignment of referential human gaze and speech production (Griffin & Bock, 2000; Van der Meulen, Meyer, & Levelt, 2001). This also enabled us to observe two types of reactive human gaze: One being elicited by robot gaze (potentially indicating joint attention), the other being utterance-mediated shifts of visual attention (to inspect mentioned objects).

In addition to the items, we constructed 48 filler videos (a set of 32 filler videos and a set of 16 videos that were item trials for a sub-study[2] such that we obtained twice as many fillers as we had items. To compensate for the relatively high proportion of anomalous items (only a third of all items was true and showed congruent or neutral gaze), a large number of fillers contained a correct statement and congruent robot gaze behavior. That is, 36 of 48 filler videos in total contained true statements (75%) and 24 were both true and congruent (50%). This results in an overall distribution of 66% true trials in the experiment. This bias towards true statements was intended to maintain the participant's trust in the competence of the robot. However, robot gaze can be considered relatively unpredictive since there were only 55.5% congruent trials overall, showing robot gaze to an object which was subsequently mentioned. This reduces the likelihood of gaze-following emerging for purely strategic reasons.

Twelve lists of stimuli each containing 72 videos were created. Each participant saw only one condition of an item and, in total, four videos in each condition. The order of the item trials was randomized for each participant individually with the constraint that between items at least one filler was shown. Additionally, fillers were randomized for each of the four sets of lists (each containing 12 lists), thus, no participants saw the same sequence of trials.

2.1.3. Procedure

An EyeLink II head-mounted eye-tracker monitored participants' eye movements on a 24-in. monitor at a temporal resolution of 500 Hz and a spatial resolution of 0.1°. Participants were seated approximately 80 cm from the screen. Viewing was binocular, although only the dominant eye was tracked. The eye-tracker was adjusted, calibrated and validated manually for each participant using a nine-point fixation stimulus. Before the experiment, participants received written instructions about the experiment procedure and task: They were asked to attend to the presented videos and judge whether or not the robot's statement in each was valid with respect to the scene. In order to provide a cover story for this task, participants were told that the robot system was being evaluated. Further, they were instructed that the robot would still make many mistakes and that participants' feedback was needed as feedback for a machine learning procedure, improving the robot system. Crucially, robot gaze was not required to perform the task nor did it change the assessment of sentence validity with respect to the scene (with the exception of only two fillers where sentence ambiguity affected validity). In contrast, participants were generally required to pay close attention to the robot's utterance as well as the scene in order to quickly complete their task. Each trial started with a fixation dot that appeared at the center of the screen. Participants were instructed to always focus on that dot so as to allow the system to perform drift correction when necessary. Then a video was played until the participant pressed a button or until an overall duration of 12 seconds was reached. The entire experiment lasted approximately 30 minutes.

2.1.4. Analysis

The presented videos were segmented into Interest Areas (IAs), i.e., each video contained regions that were labelled "anchor", "target" and "competitor", "robot head", or "distractor" as for instance the objects next to the anchor (see Fig. 3d). The temporarily ambiguous target noun "pyramid" from the example utterance was the *spoken reference* to two potential objects (*referents*) in the scene – the

[1] Mobile Robots Inc., Amherst, NH, United States; kindly provided to us through the CoSy/CogX group at DFKI, Saarbrücken (http://www.cognitivesystems.org).

[2] In this sub-study, we found consistent evidence for gaze-following behavior and the use of gaze for resolving ambiguous references. But since these results did not substantially extend the findings from Experiments 1 and 2, they are not reported here. For full details of this study, see Staudte (2010).

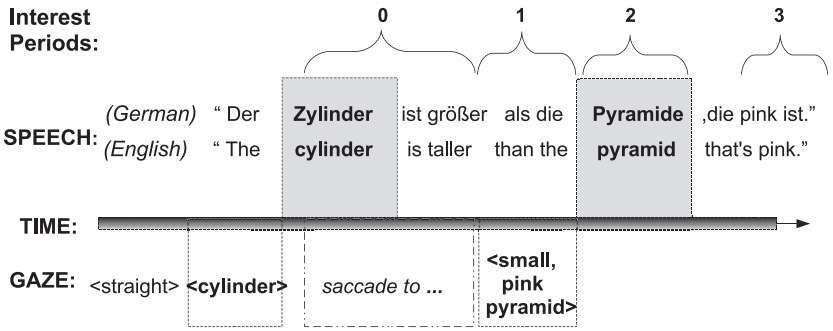


Fig. 4. The approximate timing of utterance-driven robot gaze, in a true-congruent condition.

small pink **target** pyramid or the large brown **competitor** pyramid – while referential robot gaze provided a *visual reference* to one of these objects. The small pink pyramid was considered as *target object* because the partial description "The cylinder is *taller than* the pyramid" applied to the small pink pyramid. That is, the sentence-final mention of the adjective "pink" resulted in a *correct* statement whereas mentioning the brown pyramid resulted in an *incorrect* comparison.

We segmented the speech stream into four Interest Periods (IPs) as depicted in Fig. 4. IP0 covered the whole robot gaze movement and had a mean duration of 1730 ms. IP1 stretched from the end of the gaze movement (i.e., fixation onset) to the onset of the target noun "pyramid" and had an average length of 935 ms. Thus, IP1 contained the robot's fixation towards the target object as well as some verbal content preceding the target noun ("taller than the"). IP2 stretched from target noun onset to offset. It had a mean duration of 471 ms which was constant for all conditions of an item. IP3 was defined as the 700 ms period beginning at the onset of the disambiguating color adjective. The adjective denoting the color of the referent completed the linguistic reference and unambiguously identified the actual referent. Only at that point in time was it possible to judge the Sentence Validity, which is why it is called the linguistic point of disambiguation (LPoD). Generally, responses were possible before the end of IP3 but these responses were infrequent and eventually excluded as they were either false or recognized as outliers (see below).

For the analysis of participants' fixations, all consecutive fixations within one IA and IP (i.e., before a saccade to another IA or the background occurred) were pooled and counted as one inspection. Trials that contained at least one beginning inspection towards an IA within an IP (coded as "1") are contrasted with trials that did not contain an inspection in the same slot ("0"). As a result, mean values represent inspection probabilities for a given IA/IP.

For the analysis of such un-accumulated, binary inspection data, in general, we used logistic regression (mixed-effects models with a logit link function from the *lme4* package in R; Bates, 2005). Participants and items were included as random factors, and Gaze Congruency (as well as Sentence Validity in IP3) were included as fixed factors. Chi-Square tests were applied to assess the contribution of a predictor through model reduction.[3] For comparison between levels of a factor, we further report coefficients, standard errors (*SE*) and Wald's *Z* (Baayen, Davidson, & Bates, 2008; Jaeger, 2008). For post-hoc comparisons among individual conditions in case of more than one predictor, we also use subsets of the data for each level of one predictor and fitted models with only the second predictor. *P*-values, although shown in the tables, are potentially anti-conservative (Baayen et al., 2008) so we rather refer the reader to coefficients being larger than two *SEs* for indicating significance or, additionally, generate *p*-values using Markov chain Monte Carlo (MCMC) sampling when possible (see e.g. Kliegl, Masson, & Richter, 2010, 2007; or Knoeferle & Crocker, 2009, for previous use of this method).

The elapsed time between the adjective onset and the moment of the button press was considered as the response time (RT). Trials were removed when participants had pressed the wrong button (2%). We further excluded trials as outliers when the response time was $\pm 2.5 * SE$ above or below a participant's mean (2.79%). Inferential statistics for response time are conducted using linear regression (also using mixed-effects models).

2.1.5. Predictions

Since participants had to validate the utterance with respect to a given scene, we expected participants' eye-movements to be mediated by robot speech. That is, we predicted that during sentence processing people would look at entities according to the incrementally constrained set of possible referents (Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus et al., 1995). Since the second referent was not uniquely identified until the end of a sentence (the linguistic point of disambiguation, LPoD), participants could keep several hypotheses about potential referents until then. We expected listeners' eye-movements throughout a trial to indicate which hypotheses about referent(s) were currently maintained.

We further hypothesized that people would follow not only robot speech but also robot gaze. That is, in those conditions showing referential gaze (congruent and incongruent) we expected to observe listener looks

[3] For model reduction, models were fitted by ML whereas final models are fitted using REML (see Crawley, 2007, p. 634ff).

towards the objects cued by robot gaze. In particular in IP1, when robot gaze was directed towards the target or the competitor before either had been linguistically identified, eye-movements were expected to reveal whether gaze-following occurred or not.

Our third hypothesis, mentioned in Section 1, suggested that referential speaker gaze may be considered to consist of two phases, the movement and the fixation phase, and that listeners would orient towards the direction indicated from movement onset onwards. If not only the robot fixation but indeed already the movement towards the target/competitor captured listeners' attention, the following eye movement pattern would be expected: As listeners try to figure out what the robot speaker will end up looking at, they continue to update their prediction as the camera moves across the scene such that listeners incrementally fixate what is currently in the robot's 'view'.

In IP2, we expected a continued preference to inspect the object previously identified by robot gaze. In the neutral gaze condition, however, inspections could reveal whether people used the partial utterance to constrain the domain of interpretation. That is, the target was possibly inspected more frequently than the competitor since the target – but not the competitor – was consistent with the utterance up to that point (''*The cylinder is taller than the pyramid*''). Since IP3 revealed the match (congruent condition) or mismatch (incongruent condition) of visual and linguistic references, we predicted that a match would cause listeners to continue to inspect the object they were already looking at after following robot gaze. A mismatch in referential cues was predicted to trigger an attention shift from the visual referent to the object identified by the color adjective.

Furthermore, we predicted a main effect of Gaze Congruency for response times: If participants exploited robot gaze and assumed that it indicated the robot's focus of visual attention, they would correctly anticipate the validity of statements when gaze was congruent. In contrast, when gaze was incongruent with the statement, we would predict that participants anticipate a proposition that eventually did not match the actual robot statement. Hence, a slower response time for incongruent robot gaze was expected. Since neutral gaze neither facilitated nor disrupted the judgement of the Sentence Validity, we predicted intermediate response times for this condition.

Crucially, if listeners followed and used robot gaze for purely strategic reasons, their behavior was predicted to change after a few trials when participants realized that robot gaze was almost equally often misleading as it was helping to anticipate the correct referent. Furthermore, as true statements were more frequent and expected to elicit faster response times than false statements, we also predicted a main effect of Sentence Validity on response times.

2.2. Results

2.2.1. Eye-movements

Figs. 5 and 6 show a plot of the eye-movement data for the duration of a trial and for each condition individually. The initial two seconds of a trial were preview time, the robot head started moving approximately 2000 ms after trial start. The graphs were aligned, on a trial-by-trial basis, at target noun onset (which is the beginning of IP2) to which also the robot's target gaze (fixation) was aligned during stimuli generation. The 3000 ms-time windows preceding and following this time point were divided into 250 ms-bins. Fixation proportions are computed for each IA (anchor, target, competitor, robot head, and two distractors) within each of these bins. Fixations that did not fall within an IA were counted towards background fixations and are not included in the graph. The duration of the robot gaze movement is visualized as IP0 and may be considered as an analog to human head movement.

Each plot in the time graphs shows that people initially looked mainly at the robot head. When the robot head moved towards the anchor and, more clearly, when the robot started speaking, listeners directed visual attention away from the robot's head and towards the anchor. Throughout the course of a trial, listeners rarely looked back at the robot head. The plots, however, clearly indicate gaze-following, suggesting that robot gaze was used peripherally. Gaze-following is indicated most dominantly by listeners' inspections in IP1 either on the target (in true-congruent and false-incongruent conditions) or on the competitor (true-incongruent, false-congruent), following the robot's fixation towards these objects. In contrast, in conditions true-neutral and false-neutral, neither target nor competitor were being closely attended to in IP1.

Moreover, the distractors located between the anchor and the target (called ''distractorTargetside'' in the legend) or between the anchor and the competitor (''distractorCompSide'') also received more looks when they were within the scope of robot gaze. Conditions true-congruent and false-incongruent, for instance, showed robot gaze towards the target. Consequently, in those conditions the probability to fixate the *distractorTargetside* rose along with the fixation probability of the target itself. Notably, this rise occurred already in IP0, during the robot's gaze movement phase. This suggests that listeners followed the whole gaze movement, sweeping along the scene, until the final direction was reached and the target or competitor was fixated. This way, gaze-following may be decomposed into ''movement-following'' (IP0) and ''fixation-following'' (IP1).

Further, the plots of congruent conditions show that people more frequently fixated the looked-at and mentioned object until the end of the trial while paying little attention to the other, potentially competing object. In incongruent conditions, people mostly fixated the looked-at object in IP1 and IP2 (where the referring expression is still ambiguous) and then fixated the object identified by the color adjective in IP3.

Since sentence truth did not play a role in IPs 0, 1 and 2 (because the LPoD only occurs in IP3), we collapsed each two conditions where trials were identical up to IP2 for further inspection analyses. That is, conditions true-congruent and false-incongruent were collapsed into the condition ''target gaze'', true-incongruent and false-congruent were collapsed into the condition ''competitor gaze'' and the two neutral conditions were merged to one ''neutral''-condition.

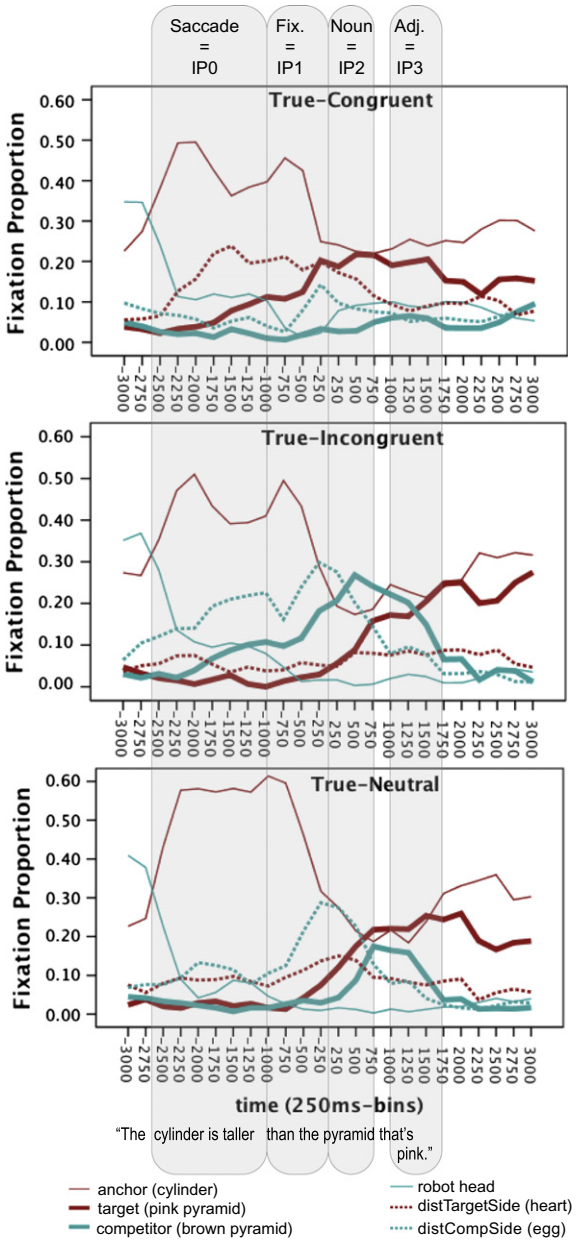


Fig. 5. Average fixation proportions in 250 ms bins across a whole trial. The plots show the **true**-sentence condition and is aligned at target noun (pyramid) onset. IP0 illustrates the average duration of the robot's gaze movement. IP1 begins with the robot's fixation towards the target/competitor and ends on noun onset. IP2 stretches from the (ambiguous) noun onset to offset, and IP3 comprises the disambiguating color adjective. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

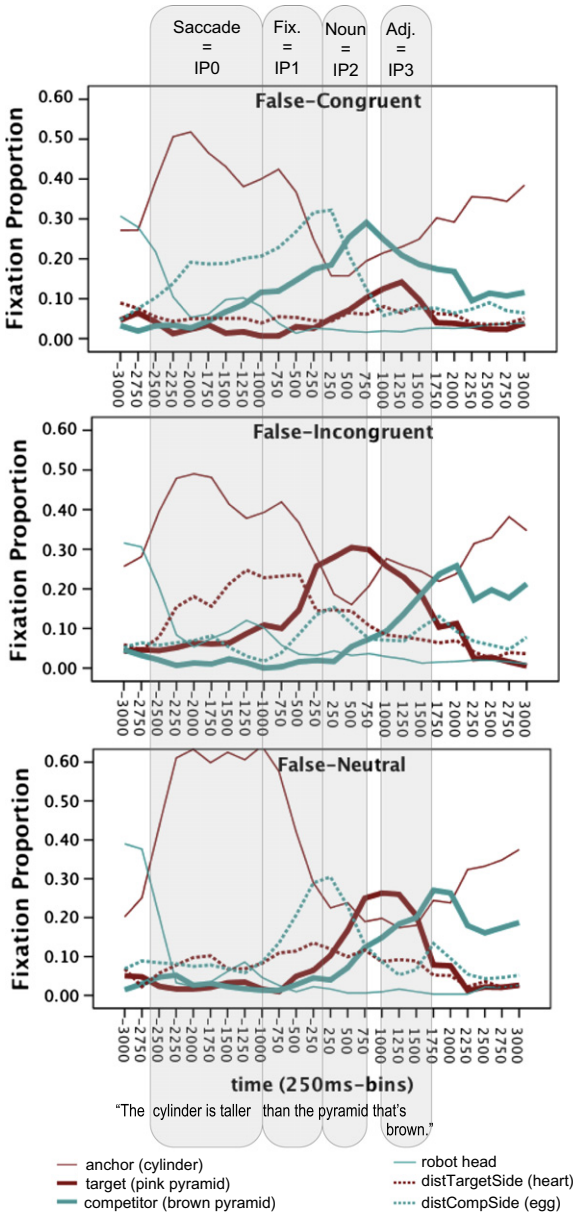


Fig. 6. These plots shows the **false**-sentence conditions and should be interpreted as in Fig. 5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the following, we summarize the inspection analyses for each IP individually, focusing on target and competitor IAs initially, before taking a closer look at the distractor IAs.

IP0 ("The CYLINDER IS TALL..."): In IP0, the robot camera swept from the anchor (cylinder) across a distractor towards the target or competitor. That is, target and competitor were not yet mentioned or reached by the robot's gaze movement and, hence, less salient than distractor objects which were located more centrally. Accordingly, a distractor object was inspected more frequently when it was within the scope of this sweeping gaze movement compared to when it was on the opposite table side (see Table 2 and Fig. 8 and further discussion of the distractor analyses below).

IP1 ("The cylinder is TALLER THAN THE"): Results from inferential statistics for target and competitor in IP1 and IP2 are given in Table 1, means are depicted in Fig. 7. During this IP, robot gaze was still the only potential cue to the intended target (e.g. big brown or small pink pyramid) and it had a main effect on people's inspection behavior (visible on the target IA: $\chi^2(2) = 146.78$; $p < 0.001$ and also the competitor IA: $\chi^2(2) = 121.65$; $p < 0.001$). The graph in

Table 1
Models fitted to separate inspection data sets (interest area = target/competitor), in IP1 and IP2. The intercept in each model represents the neutral gaze condition. *P*-values indicate the significance level of the difference between the intercept and the respective predictor level.

	Predictor	Coefficient	*SE*	Wald *Z*	*p*
IP1					
IA = target	Competitor gaze	−0.6579	0.2787	−2.360	<0.05
	Target gaze	1.6821	0.2021	8.321	<0.001
IA = competitor	Competitor gaze	1.6699	0.2220	7.522	<0.001
	Target gaze	−0.5949	0.3115	−1.910	0.056
IP2					
IA = target	Competitor gaze	−1.2799	0.2366	−5.409	<0.001
	Target gaze	0.4785	0.1857	2.577	<0.01
IA = competitor	Competitor gaze	1.5116	0.2145	7.048	<0.001
	Target gaze	−0.3726	0.2847	−1.309	0.191

Model: *IA* ∼ *Gaze* + (1|*participant*) + (1|*item*), *family* = *binomial*(*link* = "*logit*").

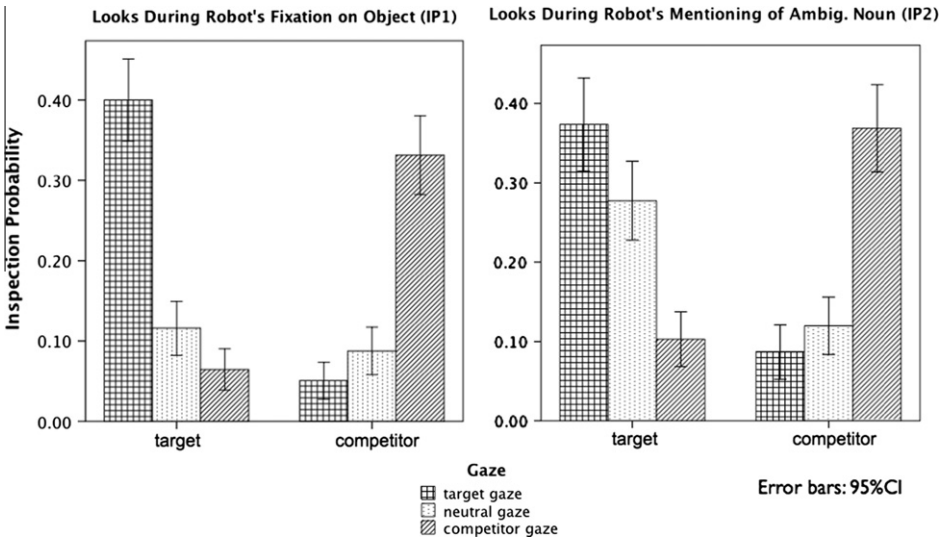


Fig. 7. Mean inspection probabilities in three gaze conditions for IP1 (left graph) and IP2 (right graph). IP1 is the 1000 ms time window preceding the target noun onset. IP2 stretches from target noun onset to offset.

Table 2
Model fitted to inspection data on distractor interest area during IP0 and IP1, with the matching condition as Intercept.

Predictor	Coefficient	*SE*	Wald *Z*	*p*
IP0				
Gaze (mismatch)	−1.7062	0.1189	−14.346	<0.001
Gaze (neutral)	0.3473	0.1116	3.113	<0.005
IP1				
Sentence (mismatch)	−0.4264	0.1518	−2.808	<0.005
Gaze (mismatch)	−1.4896	0.1664	−8.948	<0.001
Gaze (neutral)	−0.0978	0.1519	−0.644	0.519

IP0 *Model*: *Inspected* ∼ *GazeDirection* + (1|*participant*) + (1|*item*), *family* = *binomial*(*link* = "*logit*").
IP1 *Model*: *Inspected* ∼ *SentenceMatch* + *GazeDirection* + (1|*participant*) + (1|*item*), *family* = *binomial*(*link* = "*logit*").

Fig. 7 depicts these inspection probabilities and shows that people inspected the target IA with a significantly higher probability when the robot also looked at the target than when it looked at the competitor or showed neutral gaze – and the reverse holds for the competitor. When robot gaze was neutral, inspections to both IAs were equally unlikely at this point. According to previous work on sentence processing (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Sedivy et al., 1999), the mentioned comparative should constrain the domain of interpretation already at this point such that the target becomes a more likely referent than the competitor. However, this preference was not yet visible in IP1 inspections (but will be in IP2). One reason for this may be that other objects in the scene which also match the utterance so far are located more centrally (i.e., more saliently). In fact, the neutral condition clearly shows that the matching distractor (*distractorCompside*, i.e., the small, yellow egg) was indeed frequently inspected (Table 2 and Fig. 8).

IP2("The cylinder is taller than the PYRAMID"): The inspection pattern observed in IP1 persisted in IP2 for both conditions with referential robot gaze (main effect of Gaze on the target IA: $\chi^2(2) = 64.8$; $p < 0.001$ and the competitor IA: $\chi^2(2) = 87.53$; $p < 0.001$). For neutral robot gaze, participants were more likely to inspect the target IA (small, pink

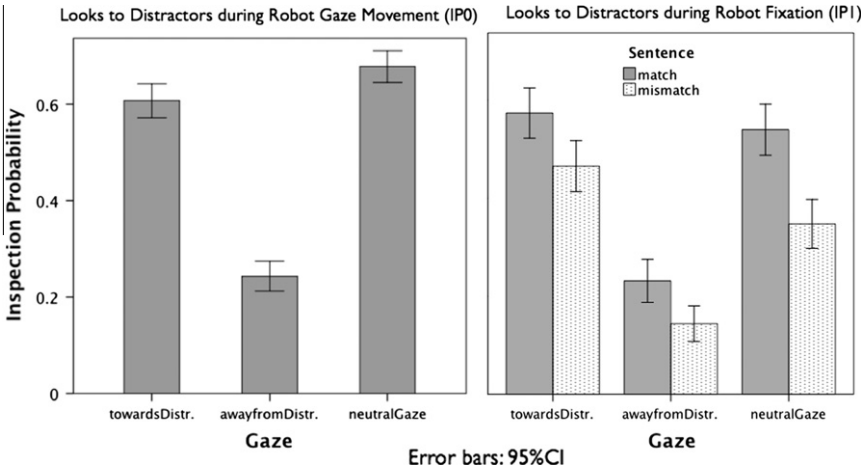


Fig. 8. Inspection proportions on distractor object in three gaze conditions (towards/away from distractor or neutral) in IP0, and additionally for both comparatives (match/mismatch with distractor size) in IP1.

pyramid) than the competitor that was consistent with the incomplete utterance so far. Pairwise comparisons between target and competitor inspections for neutral gaze showed that people inspected the target rather than the competitor ($p < 0.001$). Based on the comparative in the sentence predicate (taller/shorter), the target was the more probable referent. However, referential robot gaze introduced additional (and potentially conflicting) information since it drew attention to either target or competitor prior to IP2. Thus, when referential robot gaze was available, the preference for the object that met the linguistic constraints of the utterance (e.g., the small, pink pyramid) was no longer observable. Interestingly, participants preferred to follow robot gaze to either the target or the competitor instead.

We conducted a similar analysis (as for target and competitor) for both distractor objects located next to the anchor (see Fig. 3 again for the spatial arrangement). Since there was always one tall and one short distractor, one distractor always matched the linguistic constraints in IP1 (was shorter/taller that the cylinder) while the other did not. In the example scene in Fig. 3, the large, green heart was the *distractorTargetside* because it was between the cylinder and the small, pink target pyramid. So when the robot looked at the target, its gaze movement passed the large green heart. Figs. 5 and 6 illustrate that listeners followed robot gaze *during* the movement phase (IP0) and look at what is in its scope. Thus, each of the two distractors may be spatially more or less salient, depending on robot gaze. However, when listeners hear "The cylinder is taller than" in IP1 the small yellow egg (*distractorCompside*) would be a more probable referent than the large green heart. Thus, analyzing the distractors sheds more light onto the use of gaze, both as an early spatial cue and later, along with the linguistic constraints provided by the anchor in combination with the comparative, as a means to further constrain the domain of interpretation. There, we tested the two predictors Sentence Comparative Match (distractor size either matched or mismatched the comparative in the sentence, applicable only in IP1) and Gaze Direction Match. That is, gaze was either neutral, or the distractor was in the general direction of robot gaze (i.e., when the robot looked at the target/competitor located further away its gaze passed this distractor), or the distractor was in the opposite direction to robot gaze.

The models fitted to the inspection data on distractors during IP0 and IP1 are provided in Table 2. For IP0, both, the plots in Figs. 5 and 6 as well as Table 2, provide clear evidence for the prominence of the distractor that was in the scope of the robot's gaze movement. That is, Gaze Direction Match is a highly significant predictor ($\chi^2(2) = 363.69$; $p < 0.001$) and a distractor that matches the gaze direction, i.e., is within scope of gaze movement, is inspected more frequently than when it is located on the opposite table side (cf. the negative coefficient of the mismatching condition, or the mean inspection probabilities illustrated in Fig. 8). The plots in Figs. 5 and 6 further indicate that the distractor within gaze scope was initially, i.e., *during* the movement phase, even more salient than the target/competitor object that was finally fixated by the robot. Only later, after the robot fixation and when more linguistic material was available, the target/competitor became more salient than the respective distractor.

Table 3
Models fitted to separate inspection data sets (interest area= target/competitor), in IP3.

Predictor	Coefficient	SE	Wald Z	p
IA = target				
Validity (true)	1.5996	0.2611	6.128	< 0.001
Congr (incongruent)	0.9209	0.2679	3.437	<0.001
Congr (neutral)	1.0457	0.2630	3.977	<0.001
true:incongruent	−1.4702	0.3488	−4.215	<0.001
true:neutral	−1.1999	0.3414	−3.514	<0.001
IA = competitor				
Validity (true)	−1.6345	0.2983	−5.480	< 0.001
Congr (incongruent)	−0.2048	0.2293	−0.893	0.371
Congr (neutral)	−0.0803	0.2237	−0.359	0.719
true:incongruent	1.4373	0.3804	3.778	<0.001
true:neutral	0.4847	0.3980	1.218	0.223

Model: *IA* ~ *Sentence Validity* * *Gaze Congruency* + (1|*participant*) + (1|*item*), *family* = *binomial*(*link* = "*logit*").

As illustrated by the mean inspection proportions in Fig. 8, the direction of the robot's fixation in IP1 was also (or still, continuing from IP0) a very dominant cue that mainly determined where listeners looked. Nevertheless, we found main effects for both Gaze Direction Match ($\chi^2(2) = 195.13$; $p < 0.001$) and the Sentence Comparative Match ($\chi^2(1) = 41.81$; $p < 0.001$, see also Table 2). That is, even though listeners' visual attention was primarily influenced by robot gaze, recent linguistic information (such as the comparative) was also picked up and used to *incrementally* constrain the domain of interpretation, as reflected by inspection probabilities. This shows that listeners actively attend to both modalities – speech and gaze – on-line. Together, these findings reveal an influence of gaze *during* the movement and not only after the actual fixation.

IP3 contains the LPoD specifying which pyramid is being mentioned eventually. It is examined separately from IP1 and IP2 since both factors Sentence Validity and Gaze Congruency now affected participant behavior. Fitting and comparing multiple linear mixed-effects models (for target and competitor IAs separately) shows that both predictors and, primarily, their interaction significantly contribute to a model of the respective data set.

Firstly, we observe a robust main effect of Sentence Validity (Table 3). The positive coefficient of a predictor level (e.g., in the case of 'true') indicates a higher inspection probability for a given interest area over the intercept level 'false'. That is, listeners were more likely to inspect the linguistically identified object (which is the target in true statements and the competitor in false statements). Secondly, the interaction suggests that in congruent conditions listeners continuously inspected the object fixated and mentioned by the robot whereas in incongruent conditions visual attention was typically shifted from the object fixated by the robot to the object actually mentioned by the robot (see also the plots in Figs. 5 and 6).

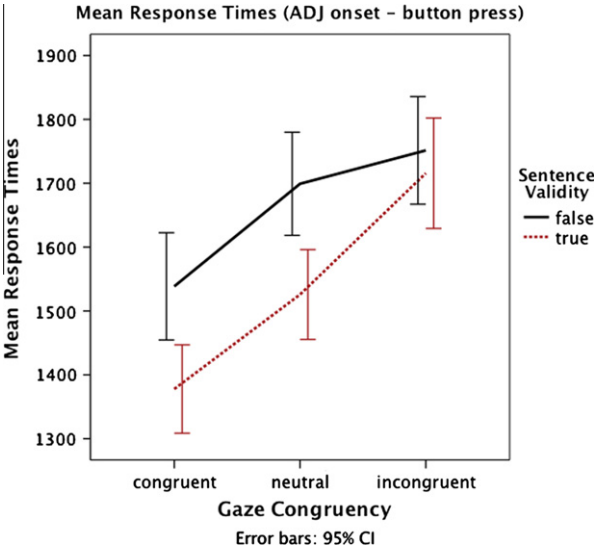


Fig. 9. Average response times for true and false statements, per Gaze Congruency condition.

Table 4
Summary of final response time model.

Predictor	Coefficient	*SE*	*t*-Value
Truth (true)	−166.874	48.744	−3.423
Congr (incongr.)	203.771	48.699	4.184
Congr (neutral)	155.575	47.504	3.275
true:incongruent	139.539	69.262	2.015
true:neutral	−4.033	68.152	−0.059

	Coefficient	MCMCmean	pMCMC	Pr(>|*t*|)
Truth (true)	−166.874	−167.243	0.0002	<0.001
Congr (incongr.)	203.771	204.294	0.0002	<0.001
Congr (neutral)	155.575	155.603	0.0008	<0.005
true:incongruent	139.538	138.640	0.0470	<0.05
true:neutral	− 4.033	−4.043	0.9574	0.953

Model: *RT* ∼ *Sentence Validity* ∗ *Gaze Congruency* + (1|*participant*) + (1|*item*).

2.2.2. Response time

Mean response times are plotted in Fig. 9 and the corresponding model is given in Table 4. Trials were excluded from response time analysis when participants gave a wrong answer (4%) or when they were considered outliers (1.69%). Model reduction on the remaining data suggests that both predictors, Sentence Validity and Gaze Congruency, contribute to fitting a model to the data (Sentence Validity: $\chi^2(1) = 19.06$; $p < 0.001$, Gaze Congruency: $\chi^2(2) = 60.43$; $p < 0.001$). Model simplification further suggests that the interaction of the two predictors is marginally significant ($\chi^2(2) = 5.598$; $p = 0.061$) but with more degrees of freedom and a higher BIC (15,897.6 versus 15,889.3 of the model without interaction) it is unclear which is the best model. We include a summary of the model containing the interaction in Table 4 along with *p*-values obtained by MCMC-sampling (a negative coefficient reveals a shorter response time of the given level compared to the intercept level). Participants were significantly faster in responding when they had to give a positive answer (true condition) than when the robot's utterance was false. Moreover, people were also significantly faster in congruent trials, that is, when the robot's gaze and utterance referred to the same object, compared to when robot gaze was neutral or incongruent. Reorganizing predictor levels within the model revealed that neutral and incongruent gaze did not differ significantly in the elicited response time, although there is a numerical tendency for increased response time in incongruent trials. The reason for the lack of significance may be related to the difference between true-neutral and false-neutral conditions: True-neutral and true-congruent behaviors similarly elicit and confirm the correct hypothesis, while false-neutral and false-incongruent conditions both initially elicit inspections to the target and then confront participants with conflicting information by identifying the competitor. It is, thus, not surprising that the difference in response times between false-neutral and false-incongruent conditions is relatively small while the difference between true-neutral and true-incongruent is relatively large. In fact, it is noteworthy that true-congruent is significantly faster than true-neutral (according to post-hoc pairwise comparison with $p < 0.01$) since

linguistic constraints select the target in both cases. It seems that robot gaze is such a strong, assuring cue that participants maintain an even stronger hypothesis about the validity of the sentence and, thus, respond faster when it is confirmed.

2.2.3. Combined analyses

We analyzed the two dependent variables, response time and inspection data, separately so far mainly because they have different properties. However, it appears reasonable to investigate the relation of these two dependent variables since they are both observed in response to the same manipulation of the stimuli. This way, one set of data can possibly help to examine reasons for the variation of another set of data. In our case, eye-movement data is observed as an on-line measure *during* exposure to the stimuli while response time is a measure recorded *after* perceiving the stimulus. Since our main manipulation concerned the robot's gaze direction (Gaze Congruency) which occurred in the middle of a trial (as opposed to Sentence Validity which is a manipulation of the final part of a trial), participants' eye movements may potentially help to understand and explain how people's visual attention during a trial relates to their response time.

Recall that we found that the predictor Gaze Congruency affected response time, but in precisely what way remained speculation. We further found that participants followed robot gaze to an object and hypothesized that this visual referent may be considered to predict the linguistic referent. To shed some light on the relation between gaze-following and the response time effect, we included people's inspection behavior during IP1 (robot gaze towards target/competitor) as a predictor for a model of response time data. We predicted that, if the early visual cue to a potential referent led listeners to form a hypothesis about upcoming linguistic references, those listeners who actually followed gaze would be faster in congruent trials. Similarly, following speaker gaze to an object that was eventually not mentioned would mislead listeners and, thus, slow them down. In contrast, ignoring speaker gaze and not looking at the visual referent was predicted to flatten this effect and result in a response pattern similar to the neutral gaze condition.

The data were coded as *following robot gaze* ('1') when listeners had inspected the IA that the robot looked at at least once during IP1 and as *not following robot gaze* ('0') otherwise. Since we were interested in the effect of gaze-following, the neutral gaze condition was dropped in this analysis. The resulting data set included participant and item information, the experimental condition (true/false, congruent/incongruent) as well as whether participants followed the robot gaze to the visual referent or not, and their response time. Model reduction showed that the predictor GazeFollowed interacts with Gaze Congruency ($\chi^2(3) = 11.425$; $p < 0.01$). The interaction introduces a larger BIC to the model but log-likelihood is largest, too, and since we are interested particularly in this interaction we include it in the final model summarized in Table 5.

Fig. 10 depicts mean response times as a function of (i) whether people followed robot gaze (represented by lines "follow" versus "NOTfollow"), (ii) whether robot gaze was congruent or not, and (iii) whether the sentence was valid or not. Crucially, the interaction between GazeFollowed and Gaze Congruency (which is also visible in Fig. 10) suggests that facilitation as well disruption effects of the gaze cue were larger when participants actually followed that cue and looked at the potential referent. Participants that did not look at the visual referent showed smaller differences in their response times. Interestingly, the main effect of Gaze Congruency – even though smaller – remained, suggesting that people did take notice of gaze and the visual target referent, though possibly covertly. These results further support the claim that speaker gaze cues a visual referent which influences listeners' hypotheses about the utterance.

2.3. Discussion

The results of Experiment 1 show that listeners follow robot gaze and that this influences the time needed to validate the utterance. The listener behavior we observed in response to robot gaze and utterance is in many respects similar to what Hanna and Brennan (2007) observed in their studies. We similarly found that: (i) Listeners begin to orient visual attention in the same direction as the robot/speaker within 1000 ms after "VPoD" (visual point of disambiguation, which corresponds to our robot's gaze onset), (ii) Listeners follow the robot/speaker's gaze during scene and utterance comprehension, (iii) Listeners use this gaze cue for early disambiguation of a spoken reference. That is, they look at the target rather than the competitor well before the LPoD.

We therefore conclude that people use robot gaze in a manner similar to how they use human gaze. Firstly, the persistence of the observed congruency effects across the experiment, in particular, seems to suggest that people *automatically* follow speaker gaze. That is, the observed fixation patterns in response to robot gaze are also consistent with and extend the idea that gaze elicits reflexive visuo-spatial orienting (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999). Specifically, these results are in line with studies showing that people reflexively follow gaze cues and also other direction-giving cues such as arrows (Langton, Watt, & Bruce, 2000; Ristic, Friesen, & Kingstone, 2002). Moreover, our results go beyond the findings from these studies in showing that reflexive visuo-spatial orienting seems to also be elicited by dynamic gaze cues and during a speech-related task. To our knowledge, this interaction of reflexive visual cueing with on-line language comprehension has not been explored previously. Secondly, the observed effects of speaker gaze congruency on utterance comprehension in terms of response time further suggest that people *use* the visual information provided by speaker gaze. That is, independent of whether gaze-following is reflexive or intentional, the visual information that people acquire through that shift in visual attention appears to be considered as an indicator to the next referent. Our findings are also the first direct quantitative evidence that using speaker gaze to mentioned objects facilitates utterance comprehension, and importantly, gaze to irrelevant objects disrupts utterance comprehension compared to a neutral condition. That

Table 5
Summary of response time model and according *p*-values from Markov chain Monte Carlo (MCMC) sampling.

Predictor	Coefficient	*SE*	*t*-Value
Truth – true	−156.69	62.62	−2.502
Congr – incongruent	140.24	63.08	2.223
GazeF – followed	−84.87	77.78	−1.091
true:incongruent	102.91	90.32	1.139
true:followed	−22.76	110.04	−0.207
incongruent:followed	205.48	108.47	1.894
true:incongruent:followed	107.88	155.18	0.695

	Coefficient	MCMCmean	pMCMC	Pr(>|*t*|)
Truth – true	−156.69	−152.06	0.0206	0.013
Congr – incongruent	140.24	144.07	0.0212	0.027
GazeF – followed	−84.87	−83.70	0.2882	0.276
true:incongruent	102.91	95.55	0.2966	0.255
true:followed	−22.76	−31.23	0.7802	0.836
incongruent:followed	205.48	201.00	0.0742	0.059
true:incongruent:followed	107.88	116.37	0.4652	0.487

Model: *RT* ∼ *Sentence Validity* ∗ *Gaze Congruency* ∗ *GazeFollowed* + (1|*participant*) + (1|*item*).

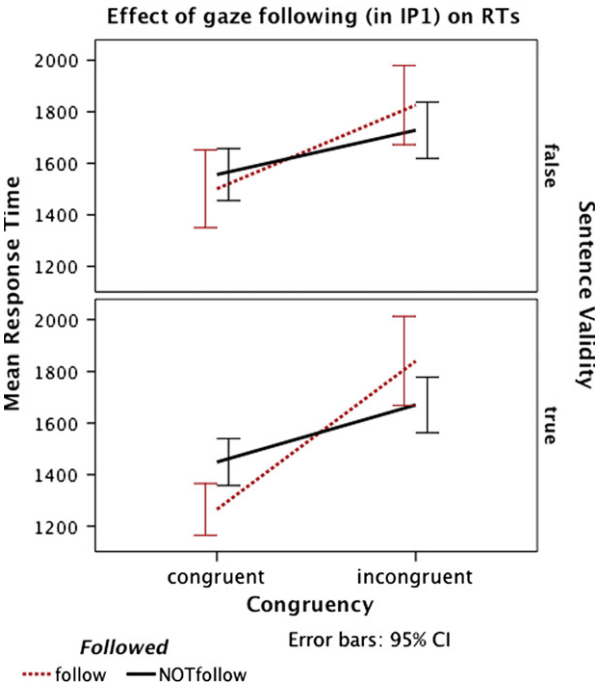


Fig. 10. Inspection pattern predicting response times. When people had followed robot gaze to the target/competitor in IP1, Gaze Congruency had a greater effect on response times.

is, there essentially is a cost to using gaze when it is not congruent with the utterance (see also Section 4).

Furthermore, we would like to point out that Hanna and Brennan reported in their studies that listeners rarely looked at the speakers' face to detect where the speaker was gazing at and rather used the speaker's head orientation peripherally. This is additional support for the claim that the type of robot gaze used in our studies – that is, as a combination of head and gaze movement – can in principle be used in much the same way that human speakers' gaze is used even though the robot has no anthropomorphic appearance and no human-like eyes. We suggest that it is sufficient for people to ascribe the function of 'seeing' to the camera in order to elicit similar behavior that human gaze elicits.

Finally, visuo-spatial orientation induced by speaker gaze seems to constrain the domain for utterance interpretation which in turn affects reference resolution. That is, listeners appear to firstly spatially constrain the domain of interpretation based on the speaker gaze movement prior to fixation. Only then, information about the more specific gaze target may be integrated with the linguistic context to determine the conceptually most plausible referent(s). While this is generally in line with Hanna and Brennan's results, their analyses and results were based on speaker fixations only (in contrast to gaze and head movement). The response time data from Experiment 1, thus, provide additional support for the (incremental) influence of both gaze movement and fixation on reference resolution and as such are a novel contribution to the investigation of how non-verbal cues like gaze precisely affect language processing. Additionally, the unpredictability of the gaze cues and the persistence of the observed effects both support the claim that gaze cues elicit reflexive orienting and it is unlikely that gaze-following is adopted as a strategy to efficient task completion. The eye-movement data from Experiment 1 further provide a novel contribution to understanding the time course of multi-modal information integration involving a coarse distinction of gaze movement and fixations as well as their interplay with the unfolding speech stream. Of course, robot gaze movement is much slower compared to human gaze (saccades) and head movement even. Nevertheless, we believe that the conceptual distinction of movement versus fixation is valid and may certainly be compared to the slower and more overt human head movement.

The presented evidence provides strong support for the hypothesis that listeners incrementally integrate speaker gaze and linguistic information with each other. On the one hand, these results lend credence to the assumption that mechanisms typically involved in visually establishing *joint attention* are also applied during this interaction with

the robot. On the other hand, this evidence confirms and quantifies the otherwise rather intuitive notion of facilitation of utterance comprehension by gaze while uncovering the incremental integration procedure during gaze and speech comprehension. The exact cause for the observed facilitation effect, however, has neither been established in the literature, nor can it be concluded from Experiment 1. A purely visual cueing effect as well as inferences of referential intentions are conceivable explanations and Experiment 2 sought to tease those apart.

3. Experiment 2

The results of Experiment 1 suggest that speaker gaze is a dominant cue which guides visual attention in an automatic fashion and that this further influences utterance comprehension. However, there are two possible explanations for the observed response time effects. Either listeners infer referential intentions from the speaker's gaze so that the expectation of a referent facilitates (or, if incorrect, disrupts) comprehension (the *Intentional Account*). Or, listeners attend to the visual referent but do not infer any referential intentions and speaker gaze simply induces a visual attention shift either to the correct object at the right time (facilitation) – or not (disruption). We call this the *Visual Account*.

A purely Visual Account would explain the facilitating/disruptive influence of speaker gaze in terms of a "bottom-up" process: A whole field of research has shown in many studies that reflexive orienting of visual attention is triggered in response to (typically static) gaze cues (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999). It is therefore reasonable to assume that robot gaze may similarly draw attention to an object while the utterance subsequently draws attention to that same (congruent) or another object (incongruent). Thus, incongruent gaze elicits an *additional* shift of visual attention before utterance comprehension is completed. This additional shift could simply add to the total time needed to comprehend and respond, thus, accounting for the observed increase in response time.

However, the effect of speaker gaze could also be explained in terms of a (mis)match in the elicited expectations about which objects were to be mentioned and the actual utterance. Previous studies on the interpretation of human gaze have revealed that gaze is an extremely versatile cue which reflects attentional states as well as mental states (Baron-Cohen, 1995; Baron-Cohen, Wheelwright, et al., 1997; Meltzoff & Brooks, 2007). We therefore hypothesize that listeners' use of speaker gaze may also be driven "top-down", that is, by the belief that gaze also reflects attentional and intentional states and thus reveals what the speaker intends to mention (Intentional Account, see also Becchio et al., 2008, for an overview of further evidence for an intentional interpretation of gaze cues). That is, participants in our experiments may have reflexively oriented towards the object that the robot gazed at, but further attributed the intention of mentioning this object to the robot. An incongruent reference therefore would entail a revision in referential expectations which would slow comprehension.

Thus, if the facilitation/disruption effect of speaker gaze on comprehension was indeed due to the inferred referential intentions, we predicted that referential speaker gaze would not only affect how fast references were resolved but also which object was believed to be the intended referent of the utterance. Such behavior would provide evidence supporting the Intentional Account. Further implications of this are that, if gaze were shown to affect beliefs about which referent the (robot) speaker *intended* to talk about, this would clearly suggest that listeners interpret (even robot) gaze with respect to attentional and intentional states, and such an inference of referential intentions would indicate that *joint attention* is possible in HRI.

Experiment 2 more thoroughly investigated how gaze affects reference resolution when participants have to correct the robot utterance. A verbal correction implicitly required listeners to identify the referent they believed was intended by the robot, thereby avoiding the need to explicitly ask listeners and request a more conscious choice. Thus, listeners were engaged in a task designed to reveal the relative importance of linguistic and gaze cues for identifying an intended referent. Participant behavior was mainly analyzed in response to false utterances which actually required a verbal correction. These were, as in Experiment 1, accompanied by either congruent, incongruent or neutral gaze. Thus, a false sentence such as "The cylinder is taller than the pyramid that is brown" would be accompanied by robot gaze to the brown pyramid (congruent), the pink pyramid (incongruent) or neutral gaze. Participants then produced a correction involving either the brown pyramid and a predicate change ("The pink pyramid is *shorter* than the brown pyramid"), or the pink pyramid and a change in referents ("is taller than the *pyramid that is pink*"), revealing what they thought was the message intended by the robot. The aim of this experiment was to determine whether and to which extent robot gaze modulates listeners' beliefs about referential intentions. Additionally this experiment served to establish whether the previously observed visual attention pattern was robust to changes in the task and could be replicated. A post-experimental questionnaire further sought to assess the general beliefs and impressions participants obtained from the interaction with the robot.

3.1. Method

3.1.1. Participants

Thirty-six native speakers of German, again mainly students enrolled at Saarland University, took part in this study. Importantly, none of them had taken part in Experiment 1. All participants reported normal or corrected-to-normal vision.

3.1.2. Materials

We used the same set of stimuli that was used for Experiment 1. That is, 24 items were used which occurred in six conditions each. The conditions resulted from the manipulation of Sentence Validity (true/false) and Gaze Congruency (congruent, incongruent, neutral; see Fig. 3). Because we wanted to mainly analyze the correction statements participants produced, false robot utterances were

of particular interest in this experiment. For the previous example sentence and scene, a false utterance such as "The cylinder is taller than the pyramid that is brown" would be accompanied either by robot gaze towards the brown pyramid (congruent), the pink pyramid (incongruent) or neutral gaze. In those utterances, there were two linguistic cues identifying the referent. The first cue was the comparative (*taller than* or *shorter than*) and the second cue was the object color. False statements were false when these two cues did not identify the same referent, e.g., when the cylinder was not *taller* than the *brown* pyramid. Thus, participants could repair such an utterance either by changing the predicate (i.e., the comparative) or by changing the referent (i.e., the color adjective) in their correction sentence.

The neutral condition provided a baseline, revealing any bias towards either repair option in the absence of gaze. When referential robot gaze was present, it emphasized one of the potential referents: Either it supported the mentioned object (identified by color) or it supported the alternative object (matching the comparative, not color). Details on the referential variation for the three *false* conditions are shown in Table 6.

3.1.3. Procedure

In this experiment, participants were instructed to give an oral correction of the robot's utterance when they thought that the robot had made a mistake. This formulation was deliberately kept rather vague so that participants were free to interpret "mistake" in a way they found appropriate. The "cover story" for this experiment remained the same as in Experiment 1, i.e., participants were told that the robot system needed to be evaluated. They were further told to start their correction with the same object reference that the robot started with, making it easier for the system to learn from the corrected sentences. Once more, this explanation served as a cover story making the task appear plausible. Participants' utterances were recorded from trial start to end, that is, from video onset until a button was pressed, indicating that the given correction was complete. Thus, the experiment was self-paced and participants could start their utterance at any time during a trial. Participants' sentences were recorded using a mobile microphone connected to an Asio AudioCard. The eye-tracker adjustment and calibration procedure as well as drift correction and presentation of the stimuli were otherwise identical to Experiment 1.

Table 6
Linguistic and visual references to objects in three congruency conditions for a false sentence, e.g., "The cylinder is taller than the pyramid that is brown" where the small pink pyramid would be considered as target. Note that the comparative "taller" points to the small target pyramid, whereas the color adjective "brown" points to the competitor pyramid.

Condition	Gaze to:	Linguistic reference to:	
		Comparative	Color
False-neutral:	–	Target	Competitor
False-congruent:	Competitor	Target	Competitor
False-incongruent:	Target	Target	Competitor

3.1.4. Analysis

For the analysis of the corrections, we annotated the produced sentences with respect to which object was described (in response to false robot utterances only, i.e., considering only the conditions shown in Table 6). The two categories assigned to responses were *Target* (object matching the comparative) and *Competitor* (object matching the color adjective). Alternative responses were found in 3.47% of the *false*-trials and were treated as missing values in the analysis. The reason for not including these as a third category was that they were conceptually not a homogeneous response category. This means that responses were rather treated as a binary (or dichotomous) dependent variable to which simple logistic regression was applied. The dependent variable was thus coding whether the target had been described in the correction sentence ('1') or not ('0'). While we consider only false utterances and removed Sentence Validity as a factor, the remaining predictor Gaze Congruency had again three levels: Congruent, neutral and incongruent. For the analysis, we used logistic regression similar to the mixed-effects models used for the eye-movement data in Experiment 1.

We again recorded participants' eye movements during trials in order to compare participant behavior in this study with the behavior observed in the previous study. The analysis of the eye-movement data was identical to Experiment 1.

3.1.5. Predictions

Under the Intentional Account, we predicted that robot gaze would not only affect how fast references were resolved (Experiment 1) but also which object was understood to be the referent. More precisely, we expected participants to describe the target and correct the color adjective, for instance, more often in the false-incongruent condition (when the robot looked at the target) than in the false-congruent or false-neutral conditions – even though the (false) utterance always identified the competitor at LPoD. Similarly, we predicted that participants would describe the competitor and change the comparative accordingly when the robot also looked at the competitor (false-congruent). If robot gaze, however, directed listeners' visual attention towards an object without contributing referential meaning (Visual Account), a significant difference in repair patterns across the three gaze-conditions would be unlikely. With respect to eye-movements, we essentially predicted that participants would follow robot gaze and speech, replicating the findings from Experiment 1 under a different task.

3.2. Results

3.2.1. Eye-movements

The findings on visual attention during this experiment indeed replicated the findings from Experiment 1. That is, listeners robustly followed the robot's gaze and speech to objects in the scene, irrespective of the different types of task they were given in Experiments 1 and 2. Inferential analyses of the respective IPs confirmed that listeners reliably followed robot gaze. In IP1, the target was inspected more frequently when the robot looked at the target, i.e.,

in condition "target gaze" combining true-congruent and false-incongruent, compared to when it looked at the competitor (Coeff. = −2.78, *SE* = 0.32, Wald *Z* = −8.64) or when gaze was neutral (Coeff. = −2.27, *SE* = 0.27, Wald *Z* = −8.33). Similarly, listeners were more likely to inspect the competitor when the robot looked at the competitor than when it looked at the target (Coeff. = −3.27, *SE* = 0.39, Wald *Z* = −8.34) or was neutral (Coeff. = −2.21, *SE* = 0.27, Wald *Z* = −8.29). The same pattern was observed for IP2, suggesting that listeners continued to inspect the object that had previously been looked at by the robot, even though the referring expression was ambiguous at that point. More precisely, in IP2, inspections on the target were more likely when the robot had fixated the target prior to noun onset, and the competitor was more often inspected when it was looked at previously by the robot.

Overall, the change from a response time task to a self-paced correction task did not seem to affect how gaze influenced listeners' visual attention. Instead, robot gaze was followed consistently in both settings – both under time pressure (Experiment 1) as well as in a self-paced setting (Experiment 2). On the one hand, the argument that listeners followed robot gaze as part of a strategy in order to better and faster fulfill the task is unlikely since robot gaze was neither generally helpful for task completion nor was there a need to respond particularly fast. On the other hand, it is highly unlikely that listeners followed gaze (and did not look at the robot head) purely for reasons of boredom or curiosity since gaze was frequently misleading and disrupting task completion. Rather, the replication of the eye-movement results supports the view that a listener attends to speaker gaze very closely and reliably, possibly even reflexively, and that she further considers it to reflect attentional states.

3.2.2. Sentence production

Since participants had to start a correction sentence with the same object as was used in the original sentence (anchor), we mainly found corrections that additionally involved either the target or the competitor object. To assess whether the robot gaze cue influenced the choice of the object involved in a correction and whether an object itself elicited preferences for including it in a description, we initially included two predictors, Described Object and Gaze Congruency, in our analyses. Model reduction suggested that both predictors contributed to fitting a model to the data since their interaction was significant ($\chi^2(4) = 58.12$; $p < 0.001$). With more degrees of freedom, the log-likelihood of the model with both predictors was also larger than in models with only one predictor while AIC and BIC of this model were smallest. A summary of the resulting model containing both predictors and the interaction is given in Table 7. Moreover, this table shows models for each response category individually indicating how well Gaze Congruency predicted in which condition the target (or the competitor) would be chosen. Since only false statements were considered in this analysis, the gaze condition *congruent* showed competitor gaze while the *incongruent* condition consequently showed target gaze.

The individual models were logistic regression models fitted to each response category (target/competitor) separately, accounting for the fact that the response categories are not independent. While the results from the inferential analyses are provided in Table 7, we also computed mean proportions of corrections involving the target/competitor and plotted them for visualization purposes in Fig. 11.

In almost 67% of their correction statements in the neutral gaze condition participants preferably gave this correction sentence: "The cylinder is *shorter* than the pyramid that is brown." That is, in the neutral baseline condition we observed a general preference to build a corrected sentence involving the competitor (which has been linguistically identified by the mentioned color in false trials), changing the comparative in the predicate accordingly. This is depicted in the central condition in Fig. 11 and confirmed by the fixed effect of predictor Described Object in Model1, Table 7. The overall preference to keep the more explicitly mentioned object (color match) remained dominant in all three gaze conditions and can most likely be explained by two facts. Firstly, gaze is frequently incongruent in our stimuli (and often considered incorrect) whereas speech is always fluent and clear. This may have induced a general bias to trust the competence of language rather than the gaze cue. Consequently, linguistic referential cues were preferred information for the identification of the intended referent (while gaze cues "only" modulated this process). Secondly, it has been shown that people prefer to use absolute (shape and color) to relative features (size, location) for the production of referring expressions (Beun & Cremers, 1998). That is, among the linguistic referential cues, color was simply the more dominant cue to an intended referent.

A positive coefficient of the predictor Gaze Congruency in Model2 and Model3 is interpreted as a larger probability of describing the according object in a given predictor level. The results in Table 7 therefore indicate that participants corrected an utterance mentioning the target (i.e., changing the reference by changing the color adjective) significantly less often when robot gaze was directed towards the competitor (false-congruent) compared to when

Table 7
Summary of the resulting model (Model1) and summaries of models for separate outcome categories (mention or not mention of target/competitor).

Predictor	Coefficient	*SE*	Wald *Z*	*p*
Object-target	−3.9763	0.364	−10.924	<0.001
Congr-incongr.	−1.4017	0.295	−4.749	<0.001
Congr-neutral	−1.1314	0.299	−3.786	<0.001
target:incongr.	3.0129	0.438	6.887	<0.001
target:neutral	2.3959	0.444	5.393	<0.001
Described object 'target' (Model2)				
Congr-incongr.	2.2738	0.395	5.763	<0.001
Congr-neutral	1.7608	0.395	4.454	<0.001
Described object 'competitor' (Model3)				
Congr-incongr.	−2.0161	0.362	−5.566	<0.001
Congr-neutral	−1.6179	0.363	−4.461	<0.001

*Model*1: *UsedInAnswer* ∼ *DescribedObject* ∗ *Gaze Congruency* + (1|*participant*) + (1|*item*), *family* = *binomial*(*link* = "*logit*").
Model2: *Target* ∼ *Gaze Congruency* + *(1|participant)* + *(1|item)*, *family* = *binomial(link = "logit")*.
Model3: *Competitor* ∼ *Gaze Congruency* + (1|*participant*) + (1|*item*), *family* = *binomial*(*link* = "*logit*").

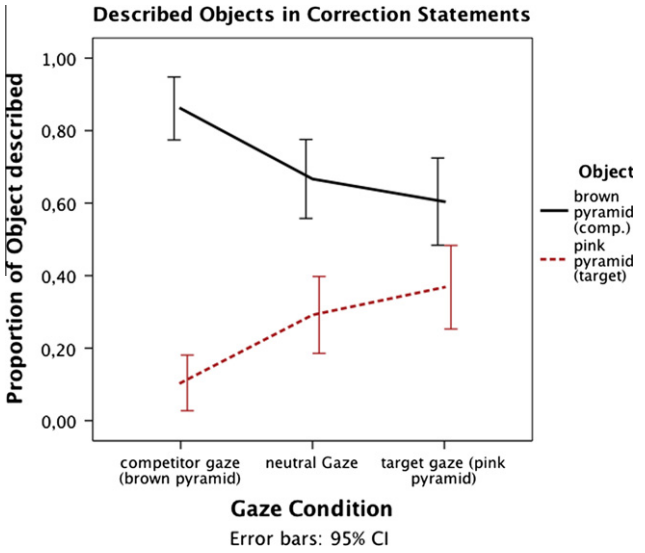


Fig. 11. Proportion of objects described in response to false robot utterances of the form "The cylinder is bigger than the pyramid that is brown".

the robot actually looked at the target (false-incongruent) or its gaze was neutral (false-neutral). These results are depicted by the dotted line in Fig. 11. Similarly, participants chose to give a scene description involving the competitor (changing the predicate, i.e., the comparative) with significantly higher probability when the robot's gaze was directed towards the competitor compared to when it was target-bound or neutral. This result is depicted by the continuous line in Fig. 11. That is, the robot's gaze increased the likelihood of correcting the competitor or target in the congruent or incongruent condition, respectively.

Another observation suggesting that gaze did affect reference resolution became apparent when analyzing corrections in response to *true* robot utterances. Although we did not expect participants to correct true statements, interestingly, we observed that in 14.6% of true-incongruent trials (i.e., in 21 trials, distributed across 14 out of 36 participants) the robot utterance was corrected with a sentence describing the competitor which the robot had looked at (true-congruent was corrected in 4 trials, true-neutral only in 1 trial). That is, participants corrected both the predicate and the reference in their response. This suggests that participants believed that the robot was indeed talking about the competitor, which it looked at, even though both the comparative and the mentioned color uniquely identified the target object. This result is surprising given that a task requiring sentence correction should induce a clear focus on the utterance. In the mentioned true-incongruent trials, however, participants most likely did not see the target otherwise they would have realized that the utterance was in fact correct. Instead, they must have focused completely on the object that the robot had fixated (competitor) leading to the unnecessary production of a correction sentence involving the competitor.

3.3. Discussion

The production results of our study suggest that robot gaze not only triggers reflexive visuospatial orienting but that it can restrict the listener's domain of interpretation. That is, listeners indeed *use* speaker gaze to infer the *intended* referents. Since in the presented study participants were asked to verbally correct the speaker's statement in a self-paced setting with no time pressure on their responses, the reflexive shift of visual attention alone cannot account for the chosen object that people describe in their corrected sentences, indicating which object they understood to be the intended referent. Instead, the results from the correction analysis support our hypothesis that robot gaze is considered to reflect its attentional state and, further, its intention to talk about an object that it looks at. These findings also suggest that listeners in fact try to establish joint attention and integrate the *visual reference* derived from gaze with their on-line interpretation of speech. Thus, these observations favor the Intentional Account for explaining the facilitation/disruption effects of referential speaker gaze on comprehension reported in previous sections.

Consequently, speaker gaze not only happens to attract listeners' visual attention to a target which is then mentioned, but it also conveys information about what the speaker presumably intended to mention next. We suggest that this is the reason why incongruent robot gaze attracted visual attention to an irrelevant object/location and disrupted utterance comprehension. In contrast, Friesen and Kingstone (1998), who employed a static and stylized gaze cue, found no disruption effect on target detection in their "un-cued" condition which similarly involved cueing of an irrelevant object/location.

Previous findings have suggested that such response behavior is not unique to human (or robot) gaze but that other attention directing cues such as arrows trigger similar reflexive behavior. However, Ristic, Wright, and Kingstone (2007) have shown that a gaze cue primes a location more reliably than arrows where the priming effect is subject to color congruency between the arrow and the actual target stimulus. According to the authors, this indicates that the attention effect for gaze is more strongly reflexive than

for arrows. An additional or alternative explanation for this reliable attention effect of eyes/gaze may be related to intentional gaze processing (Bayliss et al., 2006; Becchio et al., 2008; Castiello, 2003). Bayliss et al. (2006) have shown, for instance, that a visual referent that was looked at by another person receives higher likability scores than a not-looked at object. Another series of studies conducted by Castiello (2003) has shown, for instance, that people even infer motor intentions from an actor's gaze. Based mainly on these results, Becchio and colleagues argue that gaze potentially enriches the representation of a visual referent and they rather vaguely propose a "mechanism that allows transferring to an object the intentionality of the person who is looking at it" (Becchio et al., 2008, p. 256) which they call "intentional imposition". The findings from Experiment 2 may be considered to support the view that the robot's/speaker's intention to mention an object is transferred to it by means of gaze, and that people retrieve this referential intention when following gaze.

It is also worth mentioning that the reflexive attention shift effect as discussed above typically occurs if the target stimulus appears within a short time window after the cue (up to 300 ms for unpredicted and cued target locations and 700 ms for predicted target locations as shown by Driver et al., 1999). Our response time data from Experiment 1 clearly show that listeners are still influenced after sentence ending, that is after 2500 ms, by the previously performed robot gaze cue even though it is largely unpredictive (0.55 probability for predicting the linguistic referent).

Our data therefore provide further support for the view that gaze is indeed processed as an intentional cue as suggested by Becchio and colleagues. Moreover, our results suggest that intentional gaze processing is applied not only to human eyes but also when faced with an extremely simple realization of robot gaze (represented by a moving stereo camera).

4. General discussion

In the reported studies, we investigated how both speaker gaze and speech drive listeners' visual attention in a shared scene and how gaze is used to facilitate comprehension. A simple robot was employed as speaker to produce carefully controlled and aligned gaze and speech behavior. The results of these studies suggest that gaze facilitates reference resolution by, firstly, constraining the spatial domain of interpretation as a function of gaze movement, and secondly, by constraining the conceptual domain of interpretation in accordance with the concurrent partial utterance (Experiment 1). That is, with initially limited linguistic material and slow robot gaze movement, listeners scan what is in the scope of speaker gaze, establishing a spatial domain of interpretation. Based on the unfolding sentence and the speaker's fixation, listeners can then infer the referent within this spatial domain that they consider to be most likely intended by the speaker (Experiment 2). We further provide a first quantification of this effect on utterance comprehension by measuring response times for sentence validation in congruent, incongruent and neutral speaker gaze conditions. We observed a facilitating (congruent) but also disrupting (incongruent) influence of gaze-following on situated comprehension (Experiment 1).

Crucially, the robot used in these studies had a very simple appearance with almost no anthropomorphic features apart from the camera vaguely resembling a pair of eyes. That camera also served as head and eyes simultaneously and was the only moving part of our robot. Through this movement alone the robot appeared as actively performing. Despite this mechanistic appearance, we observed participant behavior that is very similar to what Hanna and Brennan (2007) observed. In their studies, listeners rarely looked at the speaker's face to detect what the speaker was gazing at and rather used the speaker's head orientation peripherally. We interpret this is as additional support for the claim that robot gaze as a combination of head orientation and gaze can in principle be used similarly to human speaker gaze, even though the robot has no anthropomorphic eyes or eye-movement. This is not entirely surprising, as Emery (2000) suggests that the eyes are only the first choice for interpreting an individual's direction of attention but not the only one. Instead he describes a hierarchy of cues (gaze, head, body), the use of which depends on their availability.

Related evidence for the importance of the actual camera *movement* (beyond the eye-like appearance) is provided by studies that explicitly investigated the role of motion for the assignment of goals and intentions to moving entities (e.g., Heider & Simmel, 1944). Using a simple animation which showed moving geometrical figures, Heider and Simmel (1944) found that those movements were often interpreted as one object hitting the other, as pushing or pulling actions or as leading and chasing events. That is, people interpreted movements of simple geometric shapes as goal-driven events, with one entity as agent and another as patient, and even hypothesized about *motives* for these events, suggesting that they ascribed goals and intentions to the moving objects.

This suggests that people do not rely exclusively on the anthropomorphic appearance of a face and/or eyes to elicit 'natural' reactions towards an agent, but that this can also be achieved by certain movement patterns, potentially only with appropriate scope and timing, and coupled with the assigned function of *seeing*. Whether the appearance is closer to a pair of eyes or a complete head may not play a particularly significant role here. In our case it seems that the camera *movement* which is aligned to the robot's utterance is in fact the reason why people attribute cognitive functions to it. While reflexive attention towards the robot camera may explain immediate gaze-following, we suggest that it is the attribution of cognitive functions (based on plausible motion) which ultimately explains why we observe an effect of robot gaze on reference resolution.[4]

[4] Simultaneously, we acknowledge that other visual attention capturing cues may elicit similar effects (cf. Tomlin, 1997, for arrow cues and language production). Such effects may simply rely on a link that the interpreting system establishes between the cue and a given task. If the cue is interpreted as intentional – as *intended* by the experimenter, for instance – then it may similarly influence also tasks such as language comprehension or production.

On this account, we feel there is considerable reason to believe that our results allow conclusions about general mechanisms involved in gaze processing. Moreover, our experimental setting offers several methodological advantages by combining the dynamics of situated and on-line speaker gaze (movement) with the precise control over timing and direction that stylized and static cues typically offer. Nappa, Wessel, McEldoon, Gleitman, and Trueswell (2009) similarly attempted to combine dynamic speaker gaze with precise control. They found, for instance, that speaker gaze towards a certain referent may induce a preference in children for including this referent as the grammatical subject in their description of the depicted event. Specifically, in these studies, children saw the human speaker and the depicted event scene in separate displays on a computer screen. The speaker performed an initial gaze and head movement towards one of two characters and uttered a sentence containing an unknown action verb, e.g., "He is blicking him." or "The rabbit is blicking the elephant". Crucially, an effect of speaker gaze on what children thought the speaker *meant* to say was only present in the first type of sentence, i.e., in the case of ambiguous pronominal references. This suggests that the gaze cues used in their paradigm (Nappa et al., 2009) had a somewhat weaker effect than in our paradigm, where speaker gaze modulated the production of the comparative (e.g., from "smaller" to "taller") in the case of unambiguous referring expressions. There are various possible reasons for this difference in effect strength: Due to the participants' young age, they may not yet have learned how powerful and reliable gaze is as a predictor for people's actions and utterances; Gaze in this paradigm seems to be a less situated and embodied cue due to the separated displays of the speaker and the depicted event; The lack of more natural gaze movement, i.e., towards the second referent and/or back towards the child, could have also reduced the impact of gaze. Further, gaze and speech cues were never concurrent in these studies so that on-line integration was not required and could not be investigated.

Hanna and Brennan's studies, in contrast, featured a real-time face-to-face interaction between listener and speaker. They further focused on the listeners' flexibility in interpreting gaze direction on-line by forcing them to re-map speakers' gaze to their own (different) object arrangement. While their aim was to investigate whether and how flexible a gaze cue is, our studies focus on examining the integration process of referential information provided by gaze and speech, especially in cases of mismatch. By using a robot as interlocutor, we created plausible mismatching references by introducing wrong or erroneous (i.e., incongruent) robot behavior. In such cases of error, re-mapping of perceived gaze was not appropriate and did not help in combining cues to one consistent reference. Instead, people had to make sense of the information they perceived by actively weighing one cue (or one modality, i.e., speech versus gaze/vision) higher than the other and eventually make a decision based on that.

One might argue that the kind of error produced in the reported studies does not occur in human–human interaction and that it therefore does not contribute to understanding human use of gaze. However, it is perfectly possible that a speaker's look and mention of an object are incoherent, for instance, when her spatial memory of the objects location is wrong or when she produces an incorrect term for an object. Therefore, we consider the proposed human–robot interaction scenario as suitable to generally provide insights about the utility of speaker gaze as an additional visual cue that relates to language and the shared environment.

Thus, the findings put forward here contribute to the understanding of how gaze cues are integrated with an utterance such that a listener may anticipate and ground a speaker's next referring expression in the shared environment. Moreover, we provide additional evidence for an Intentional Account of this effect, drawing on both initial reflexive orienting and the inference of referential intentions. The observed eye-movement patterns further shed light on the dynamics of this integration process and provide a novel way to explore the processing of gaze (or head) movement *and* the resulting fixations during simultaneous language comprehension. Even though the dynamics observable in our setting clearly differ from how people process gaze (which features extremely fast saccades and possibly less overt fixations than the robot), these findings constitute a first step towards understanding the effects of a non-static attention-directing cue, from movement onset (eye-movement but also the much slower head movement, for instance) to its final position. Crucially, the extremely overt and exaggerated gaze/head movement of our robot speaker may even have been interpreted as a gesture *intended* to direct listeners' attention. In that case, the behavior we observed would indicate that listeners have even tried to engage in *shared* rather than only *joint* attention. Independent of such an interpretation by the listener, however, the observed facilitation/disruption and integration effects discussed throughout this paper apply to (robot) speaker gaze and utterance comprehension.

5. Conclusions

Two eye-tracking experiments were presented that investigated how listeners interpret referential gaze and speech, examining whether such (robot) gaze is used to establish joint attention and to draw inferences about the intended referents. The findings from Experiment 1 reveal that listeners robustly follow speaker gaze and that they use speaker gaze (from movement onset to final fixation) to anticipate an upcoming referent such that congruent robot gaze facilitates comprehension while incongruent gaze disrupts comprehension relative to neutral gaze. We hypothesized that, while gaze-following may be reflexive initially (Visual Account), the utility of gaze may indeed rely on the attentional and intentional states that people ascribe to the (robot) speaker (Intentional Account). That is, we argued that listeners may indeed try to establish joint attention with the robot, interpreting its gaze to indicate what the robot attends to and what it intends to mention. The results from Experiment 2 support this hypothesis and show that speaker gaze modulates which object – in the case of wrong utterances – is considered as intended referent.

Thus, we have shown that people interpret robot gaze in a way that is similar to how people use and interpret human gaze. This is evidence of the utility of our experimental paradigm for investigating not only the role of robot gaze but also aspects of gaze processing in general. Accordingly, we offer insights on the cause of the comprehension benefits attributed to gaze cues as well as insights on the specific processing dynamics involved in following gaze during its initial movement phase as well as the subsequent fixation phase.

Acknowledgements

The authors gratefully acknowledge the support of the German Research Foundation (DFG) through a PhD scholarship (IRTG-715) to the first author, and the Cluster of Excellence "Multimodal Computing and Interaction". We also thank Geert-Jan Kruijff and the CoSy project for providing access to the robot. Finally, we would like to thank the three anonymous reviewers for their constructive and insightful comments.

References

Adams, R. B., & Kleck, R. E. (2003). Perceived gaze direction and the processing of facial displays of emotions. *Psychological Science, 14*, 644–647.

Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Altmann, G., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 347–386). NY: Psychology Press.

Argyle, M., & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry, 28*, 289–304.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science, 15*, 415–419.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MA: MIT Press/Bradford Books.

Baron-Cohen, S., Baldwin, D., & Crowson, M. (1997). Do children with autism use the speaker's direction of gaze strategy to crack the code of language? *Child Development, 68*, 48–57.

Baron-Cohen, S., Campbell, R., Karmiloff-Smith, A., Grant, J., & Walker, J. (1995). Are children with autism blind to the mentalistic significance of the eyes? *British Journal of Developmental Psychology, 13*, 379–398.

Baron-Cohen, S., Wheelwright, S., & Jolliffe, T. (1997). Is there a language of the eyes? Evidence from normal adults, and adults with autism or Asperger syndrom. *Visual Cognition, 4*, 311–331.

Bates, D. (2005). Fitting linear mixed models in R. *R News, 5*, 27–30.

Bayliss, A., Paul, M., Cannon, P., & Tipper, S. (2006). Gaze cueing and affective judgments of objects: I like what you look at. *Psychonomic Bulletin & Review, 13*, 1061–1066.

Becchio, C., Bertone, C., & Castiello, U. (2008). How the gaze of others influences object processing. *Trends in Cognitive Science, 12*, 254–258.

Beun, R., & Cremers, A. (1998). Object reference in a shared domain of conversation. *Pragmatics and Cognition, 6*, 111–142.

Breazeal, C., Kidd, C., Thomaz, A., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human–robot teamwork. In *Proceedings of IEEE/RSJ international conference on intelligent robots and systems (IROS'05)* (pp. 708–713).

Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97*, 523–547.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., et al. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the ACM conference on human factors in computing systems (CHI'99)* (pp. 520–527).

Cassell, J., & Thórisson, K. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence, 13*, 519–538.

Cassell, J., Torres, O., & Prevost, S. (1999). Turn taking vs. discourse structure: How best to model multimodal conversation. In Y. Wilks (Ed.), *Machine Conversations* (pp. 143–154). Kluwer.

Castiello, U. (2003). Understanding other people's actions: Intention and attention. *Journal of Experimental Psychology, 29*, 416–430.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*, 62–81.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84–107.

Crawley, M. J. (2007). *The R book*. Wiley Publishing.

Crocker, M. W., Knoeferle, P., & Mayberry, M. (2010). Situated sentence processing: The coordinated interplay account and a neurobehavioral model. *Brain and Language, 112*, 189–201.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Dovidio, J. F., & Ellyson, S. L. (1982). Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly, 45*, 106–113.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition, 6*, 509–540.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*, 283–292.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24*, 409–436.

Emery, N. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews, 24*, 581–604.

Flom, R., Lee, K., & Muir, D. (Eds.). (2007). *Gaze-following: Its development and significance*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Friesen, C., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review, 5*, 490–495.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language, 57*, 544–569.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition, 82*, B1–B14.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*, 274–279.

Hanna, J., & Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*, 596–615.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243–259.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*, 498–504.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446 (special issue: emerging data analysis).

Kanda, T., Ishiguro, H., & Ishida, T. (2001). Psychological analysis on human–robot interaction. In *Proceedings of IEEE international conference on robotics and automation (ICRA'01)* (pp. 4166–4173).

Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies, 7*, 135–169.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22–63.

Kiesler, S., Powers, A., Fussell, S., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition, 26*, 169–181.

Kliegl, R., Masson, M. E., & Richter, E. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition, 18*, 655–681.

Kliegl, R., Risse, S., & Laubrock, J. (2007). Preview benefit and parafoveal-on-foveal effects from word *n* +2. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1250–1255.

Knoeferle, P., & Crocker, M. W. (2009). Constituent order and semantic parallelism in online comprehension: Eye-tracking evidence from German. *The Quarterly Journal of Experimental Psychology, 62*, 2338–2371.

Knoeferle, P., Crocker, M. W., Pickering, M., & Scheepers, C. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition, 95*, 95–127.

Langton, S. R., & Bruce, V. (1999). Reflexive visual orienting in response to the social attention of others. *Visual Cognition, 6*, 541–567.

Langton, S. R., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Science, 4*, 50–59.

Meltzoff, A. N., & Brooks, R. (2007). Eyes wide shut: The importance of eyes in infant gaze-following and understanding of other minds. In R. Flom, K. Lee, & D. Muir (Eds.), *Gaze-following. Its development and significance* (pp. 217–241). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. (2010). Social robots are psychological agents for infants: A test of gaze following. *Neural Networks, 23*, 966–972 (Social cognition: From babies to robots).

Meyer, A., Sleiderink, A., & Levelt, W. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition, 66*, B25–B33.

Moore, C., & Dunham, P. (Eds.). (1995). *Joint attention: Its origins and role in development*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Mutlu, B., Hodgins, J., & Forlizzi, J. (2006). A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Proceedings of the 2006 IEEE-RAS international conference on humanoid robots (HUMANOIDS'06)*. Genova, Italy.

Nappa, R., Wessel, A., McEldoon, K. L., Gleitman, L. R., & Trueswell, J. C. (2009). Use of speaker's gaze and syntax in verb learning. *Language Learning and Development, 5*, 203–234.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*, 81–103.

Ristic, J., Friesen, C. K., & Kingstone, A. (2002). Are eyes special? It depends on how you look at it. *Psychonomic Bulletin & Review, 9*, 507–513.

Ristic, J., Wright, A., & Kingstone, A. (2007). Attentional control and reflexive orienting to gaze and arrow cues. *Psychonomic Bulletin & Review, 14*, 964–969.

Schroeder, M., & Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology, 6*, 365–377.

Sedivy, . J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition, 71*, 109–147.

Sidner, C. L., Lee, C., Kidd, C., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence, 166*, 140–164.

Staudte, M. (2010). Joint Attention in Spoken Human–Robot Interaction. Ph.D. Thesis Saarland University Germany. <http://scidok.sulb.uni-saarland.de/volltexte/2010/3242/>.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632–1634.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science, 10*, 121–125.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*, 675–691.

Tomlin, R. S. (1997). Mapping conceptual representations into linguistic representations: The role of attention in grammar. In J. Nuyts & E. Pederson (Eds.), *Language and conceptualization* (pp. 162–189). Cambridge: Cambridge University Press.

Van der Meulen, F. F., Meyer, A. S., & Levelt, W. J. M. (2001). Eye movements during the production of nouns and pronouns. *Memory & Cognition, 29*, 512–521.

Vecera, S., & Rizzo, M. (2006). Eye gaze does not produce reflexive shifts of attention: Evidence from frontal-lobe damage. *Neuropsychologia, 44*, 150–159.

Woods, S., Walters, M., Koay, K.L., & Dautenhahn, K. (2006). Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach. In *Proceedings of the 9th international workshop on advanced motion control (AMC'06)*.